TRANSACTIONS

OF THE

AMERICAN MATHEMATICAL SOCIETY

Coden: TAMTAM Whole No. 481 Pages 1–358
Volume 208 July 1975

PUBLISHED BY THE AMERICAN MATHEMATICAL SOCIETY
PROVIDENCE, RHODE ISLAND
1975

Transactions of the American Mathematical Society

THIS JOURNAL is devoted entirely to research in pure and applied mathematics, and includes, in general, longer papers than those in the PROCEEDINGS.

PREPARATION OF THE MANUSCRIPT. All papers should be typewritten, double-spaced, and the author should keep a complete copy.

FORM OF MANUSCRIPT. The first page should consist of a *descriptive title*, followed by an *abstract* which summarizes the article in language suitable for workers in the general field (algebra, analysis, etc.). The *descriptive title* should be short, but informative; useless or vague phrases such as "Some remarks about" or "concerning" should be avoided. Also avoid proper names unless mathematical usage associates them with the work. The *abstract* should be at least one complete sentence, and at most 300 words, with the upper limit primarily for longer papers. Included with the footnotes to your paper, but placed before the first footnote, there should be first the *AMS (MOS) subject classification numbers* representing the primary and secondary subjects of the article, which may be followed by a list of *key words and phrases* describing the subject matter of the article and taken from it. The AMS (MOS) Subject Classification Scheme (1970) with instructions for its use can be found as an appendix to Mathematical Reviews, Index to Volume 39 (June 1970). See the June 1970 Notices for more details, as well as illustrative examples.

SUBMISSION OF MANUSCRIPT. See the inside back cover of this journal.

GALLEY PROOF. When a paper with more than one author has been accepted for publication, only one set of galley proof will be sent. Joint authors should, therefore, indicate on the original manuscript which of them should receive galley proof in the event that the manuscript is accepted for publication.

BACKLOG. 300 pages. Two-thirds of the papers currently being accepted by the editors will be published in 12–14 months.

SUBSCRIPTION INFORMATION. Subscription prices for Volumes 201–214 (1975) are list \$210.00, member \$105.00. Subscription prices for Volumes 215–224 (1976) are list \$210.00, member \$105.00.

Back number prices *per volume* for Volumes 78–134 (1955–1968) are list \$18.00, member \$13.50; Volumes 135–200 (1969–1974), list \$30.00, member \$22.50. Volumes 201–214 (1975), when sold at back number prices, will be list \$30.00 per volume, member \$22.50 per volume. Volumes 1–77 (1900–1954) may be ordered through Johnson Reprint Corporation, 111 Fifth Avenue, New York, New York 10003. Volumes 1–200(1900–1974) are available on 35 mm microfilm from University Microfilms-Xerox, 300 North Zeeb Road, Ann Arbor, Michigan 48106.

Memoirs of the American Mathematical Society

This bimonthly journal is devoted to research in pure and applied mathematics of much the same nature as appears in TRANSACTIONS. An issue consists of one or more separately bound research tracts for which the author(s) has provided reproduction copy. Prior to 1975 this was published as a monograph series. The editorial committee is identical with that for the TRANSACTIONS so that papers intended for publication in this series should be addressed to one of the editors listed on the inside back cover.

Published monthly except two in April and October by the American Mathematical Society. Subscriptions and orders for publications of the American Mathematical Society should be addressed to American Mathematical Society, P.O. Box 1571, Annex Station, Providence, Rhode Island 02901. *All orders must be accompanied by payment.* Other correspondence should be addressed to P.O. Box 6248, Providence, Rhode Island 02940.

Second-class postage paid at Providence, Rhode Island, and additional mailing offices.

TRANSACTIONS

OF THE

AMERICAN MATHEMATICAL SOCIETY

VOLUME 208
July 1975

PUBLISHED BY THE AMERICAN MATHEMATICAL SOCIETY
PROVIDENCE, RHODE ISLAND

TABLE OF CONTENTS

Vol. 208 **Whole No. 481** **1975**

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

WEIGHTED SHIFTS AND COVARIANCE ALGEBRAS([1])

BY

DONAL P. O'DONOVAN

ABSTRACT. The C^*-algebras generated by bilateral and unilateral shifts are studied in terms of certain covariance algebras. This enables one to obtain an answer to the question of when such shifts are G.C.R., or not, or even when they are N.G.C.R.. In addition these shifts are classified to within algebraic equivalence.

Introduction. This paper is concerned with certain types of bounded linear operators on separable Hilbert spaces. The types are the weighted shifts, both bilateral and unilateral. These operators have been studied quite extensively and have been found to contain examples of many different types of operator behaviour [4], [15], [17]. Among other results, necessary and sufficient conditions are given here for when such shifts are G.C.R. or type I (§3.4), for when the C^*-algebra that they generate contains the compact operators (§2.5, §3.2), and for when two shifts are algebraically equivalent (§2.4, §3.3). In order to answer these questions, it is necessary to obtain a useful description of the C^*-algebras they generate and of their irreducible representations. For this purpose covariance algebras are most appropriate [3], [9], [10], [23], [24].

In the first part of this paper the results on covariance algebras that are needed are presented. Many of these results are known; they appear chiefly in [24]. In the case of the group Z, some of the proofs are conceptually easier and it seemed worthwhile to present them. The principal new result here is Theorem 1.2.1, in which it is shown that a necessary and sufficient condition on a homeomorphism ϕ of a compact space X in order that every ideal in the covariance algebra $C^*(X, \phi)$ contain an element of $C(X)$, is that the periodic points be a "small" set. The theorem is in fact proven for a general $C^*(\mathfrak{A}, Z)$.

The C^*-algebras generated by weighted shifts with closed range are completely characterized in §§2.2 and 3.1 in terms of covariance algebras $C^*(X, \phi)$.

Received by the editors July 3, 1974.

AMS (MOS) subject classifications (1970). Primary 46L05; Secondary 46K10, 47B99, 47C10.

([1]) Much of the following material is contained in a dissertation written under the direction of W. B. Arveson and presented in partial fulfillment of the requirements of the Ph. D. degree at the University of California at Berkeley.

It is shown that the space X has a certain canonical form which for a given shift makes explicit all of its irreducible representations as weighted shifts. Also this canonical form of X classifies shifts to algebraic equivalence (§2.4, §3.3). This last term was introduced by W. B. Arveson [2] to describe two operators T and S for which the map $T \to S$ extends to a *-isomorphism of $C^*(T)$ onto $C^*(S)$. For normal operators this means they have the same spectrum, so for weighted shifts we have an "induced" version of this result.

If the above remarks seem to make little distinction between unilateral and bilateral shifts, this is because as is seen in Parts II and III, the differences are much less than might have been expected. In fact the type of analysis carried out here is almost equally applicable to all classes of centered operators [21].

The terminology and notation used are the standard ones [2], [8]. Thus, for example, $L(H)$ and $C(H)$ denote the bounded linear operators and compact linear operators on a Hilbert space H, $C^*(\{\ \})$ denotes the C^*-algebra generated by $\{\ \}$ and 1, $\mathfrak{A}'$ denotes the commutant of an algebra $\mathfrak{A}$, and H_π denotes the Hilbert space on which a representation π of some C^*-algebra acts. The order of presentation is:

PART I. COVARIANCE ALGEBRAS

1.1. Representations. If χ is a *-automorphism of a C^*-algebra $\mathfrak{A}$, the semidirect product or covariance algebra $C^*(\mathfrak{A}, Z)$ is constructed as follows: Let $l^1(\mathfrak{A}, Z)$ be the set of all $\mathfrak{A}$-valued functions F on Z for which the norm

$\|F\|_1 = \Sigma_{n=-\infty}^{\infty}\|F(n)\|$ is finite. $l^1(\mathfrak{A}, Z)$ is a Banach space in this norm and if a multiplication and an involution are defined by

$$(F_1 * F_2)(n) = \sum_k F_1(k)\chi^k(F_2(n-k))$$

and

$$F^*(n) = \chi^n(F(-n)^*),$$

then $l^1(\mathfrak{A}, Z)$ becomes a Banach algebra with approximate identity. Now $C^*(\mathfrak{A}, Z)$ is defined to be the enveloping C^*-algebra [8]. Thus, for $F \in l^1(\mathfrak{A}, Z)$ put

$$\|F\| = \sup_\pi \|\pi(F)\|,$$

where π ranges over all irreducible $*$-representations of $l^1(\mathfrak{A}, Z)$. One can show [9], [24] that $\|F\| = 0 \Rightarrow F = 0$ so $C^*(\mathfrak{A}, Z)$ is defined to be the completion of $l^1(\mathfrak{A}, Z)$ in this norm. More generally, if G is any locally compact group of $*$-automorphisms of $\mathfrak{A}$, then $C^*(\mathfrak{A}, G)$ can be constructed [9], [23], [24]. To every representation ρ of a covariance algebra $C^*(\mathfrak{A}, Z)$ corresponds a pair (π, U), where π is a representation of $\mathfrak{A}$, and U is a unitary operator on H_π with the property that $U\pi(A)U^{-1} = \pi(\chi(A))$ for all A in $\mathfrak{A}$. In fact if $F \in l^1(\mathfrak{A}, Z)$, then

$$\rho(F) = \sum_{n=-\infty}^{\infty} \pi(F(n))U^n. \tag{1.1}$$

We shall express this relationship by writing $\rho = (\pi, U)$.

Let $\hat{A}$ and $\widehat{A}$ denote the spectrum, or dual, and the quasi-spectrum of $\mathfrak{A}$ respectively [8]. $\hat{A}$ can be naturally embedded in $\widehat{A}$, and $\widehat{A}$ can be endowed with the Mackey Borel structure and the Jacobson topology [8], [13].

Any representation π of $\mathfrak{A}$ has a central decomposition, $\pi = \int_{\widehat{A}}^{\oplus} c(x)\,d\mu(x)$ where μ is a standard Borel measure on $\widehat{A}$, $c(x)$ is a measurable cross section of the quotient map $\mathrm{Fac}(\mathfrak{A}) \to \widehat{A}$, and the center of $\pi(\mathfrak{A})$ consists of the diagonalisable operators M. These are the operators M_f (where f is in $\mathfrak{B}(\widehat{A})$, the bounded Borel functions on $\widehat{A}$), defined on F in H_π by $(M_f F)(x) = f(x) \cdot F(x)$ [8].

Any $*$-automorphism χ of $\mathfrak{A}$ induces an obvious map ϕ of $\widehat{A}$ into $\widehat{A}$ which leaves $\hat{A}$ invariant. Further, it is immediate from their definitions that ϕ is an isomorphism for the Borel structure and a homeomorphism for the topology.

For a representation $\rho = (\pi, U)$ of $C^*(\mathfrak{A}, Z)$, let θ denote the inner automorphism of $\pi(\mathfrak{A})$ given by $A \to UAU^{-1}$. Then the center of $\pi(\mathfrak{A})'$, the commutant of $\pi(\mathfrak{A})$, is invariant under θ. As is shown in [13], this leads to the fact that

$$UM_fU^{-1} = M_{f\circ\phi}, \tag{1.2}$$

for each $M_f \in M$. This implies that μ is quasi-invariant with respect to ϕ i.e. $\{\phi^n \circ \mu\}_{n \in Z}$ are pairwise absolutely continuous, where $(\phi \circ \mu)(E) = \mu(\phi(E))$. So if $h = d(\phi \circ \mu)/d\mu$ is the Radon-Nikodym derivative, then defining U_ϕ on $F \in H_\pi$ by

$$(U_\phi F)(x) = \sqrt{h(x)}F(\phi(x)),$$

U_ϕ is unitary and has the property that $U_\phi M_f U_\phi^{-1} = M_{f \circ \phi}$. Thus $UU_\phi^{-1} \in M'$, so $U = B \cdot U_\phi$, where B is a decomposable operator.

If $\rho = (\pi, U)$ is an irreducible representation, then it follows immediately from 1.2 (since $\pi(\mathfrak{A})' = M$), that ϕ must be ergodic with respect to μ, i.e. if $E \subset A$ is measurable and $\phi(E) = E$, then $\mu(E) = 0$ or 1. Since $\phi(\hat{A}) = \hat{A}$, μ_π is based on either $\hat{A}$ or $\widehat{\mathfrak{A}} - \hat{A}$. The former is a necessary and sufficient condition that π be a type I representation [8, Proposition 8.4.8]. If so, then π has a unique decomposition $\pi = \pi_\infty \oplus \pi_1 \oplus \pi_2 \oplus \ldots$, where π_i is a representation of multiplicity i, $1 \leq i \leq \chi_0$. In fact $\pi_i = \int_{B_i}^{\oplus} c(x)\, d\mu(x)$, where $B_i = \{x\colon c(x)$ is quasi-equivalent to some $i \cdot \nu$ with ν irreducible$\}$. Each B_i is clearly ϕ-invariant, so again by ergodicity if π is of type I, then π has uniform multiplicity. In fact, one can go further and conclude that μ is concentrated on some $\hat{A}_{n,i} = \{x\colon c(x) = i \cdot \nu$ and $\dim H_\nu = n\}$.

An ergodic quasi-invariant measure μ may have $\mu(0) = 1$ for some orbit 0 of ϕ. In this case the measure is said to be transitive. Otherwise it is called intransitive [20].

Suppose $\rho = (\pi, U)$ is an irreducible representation of $C^*(\mathfrak{A}, Z)$ for which μ_π is transitive, here this means purely atomic, based on the orbit of a_μ in $\widehat{\mathfrak{A}}$ say. If a_μ is not in $\hat{A}$, then one sees readily that ρ is not in fact irreducible. So a_μ is in $\hat{A}$, and then we have seen that $\pi = i \cdot \int_{\hat{A}_n} c(x) \cdot d\mu_\pi(x)$, and π is independent of the particular cross section $c(x)$ chosen [8]. If μ_π is purely atomic, then any cross section is measurable, and one can be chosen so that $U_\phi \pi(A) U_\phi^{-1} = \pi \circ \phi(A)$ for all $A \in \mathfrak{A}$. Then $U = BU_\phi$, where $B \in \pi(\mathfrak{A})'$. Further for such μ_π one sees that if π is not multiplicity free, then ρ is reducible.

To summarize, if $\rho = (\pi, U)$ is any irreducible representation of $C^*(\mathfrak{A}, Z)$ for which μ_π is transitive, then μ_π is based on $\hat{A}$, π is a multiplicity free representation, and $U = M_h \cdot U_\phi$, for some $h \in B(\hat{A})$, with $|h| \equiv 1$. If the point a_μ is not periodic, then the family of representations $\rho = (\pi, M_h U_\phi)$ are all possible and all unitarily equivalent. If a_μ is periodic of period k, then again the representations $(\pi, M_h U)$ are all possible and two such are unitarily equivalent if and only if

$$h_1(a_\mu) \cdot h_1(\phi(a_\mu)) \cdots h_1(\phi^{k-1}(a_\mu)) = h_2(a_\mu) \cdot h_2(\phi(a_\mu)) \cdots h_2(\phi^{k-1}(a_\mu)).$$

In general if μ_π is not transitive, then it appears that the relationship

$U_\phi M_f U_{\phi^{-1}} = U_{f\circ\phi}$ cannot always be lifted to give $U_\phi \pi U_\phi^{-1} = \pi \circ \chi$. However for the case of $\mathfrak{A}$ commutative this is of course automatic, so one has in this case that for every quasi-invariant ergodic measure μ on $\hat{A}$, $C^*(\mathfrak{A}, Z)$ has an irreducible representation on $L^2(\hat{A}, \mu)$.

While not much can be said about the intransitive irreducible representations, it is at least clear that one of them cannot be unitarily equivalent to a transitive representation. For if $V: (\pi_1, U_1) \to (\pi_2, U_2)$, then V must implement a unitary equivalence between π_1 and π_2 and thence between the centers of $\pi_1(\mathfrak{A})'$ and $\pi_2(\mathfrak{A})'$. This is clearly impossible if μ_1 is atomic and μ_2 is not.

1.2. Ideals. We now want to consider ideals in $C^*(\mathfrak{A}, Z)$. Since, as was remarked earlier, $l^1(\mathfrak{A}, Z) \subset C^*(\mathfrak{A}, Z)$, there is a natural injection i of $\mathfrak{A}$ into $C^*(\mathfrak{A}, Z)$, with $i(A)(n) = A\delta_{0,n}$, for $A \in \mathfrak{A}$. The specific question to be answered is: If I is an arbitrary nonempty selfadjoint ideal in $C^*(\mathfrak{A}, Z)$, under what conditions on $\mathfrak{A}$ and the action of Z on it can it be concluded that $I \cap \mathfrak{A} \neq \{0\}$?

If the induced action ϕ on $\hat{A}$ is free, i.e. no periodic points, then it follows from [24] that the above is true. We shall show the following. Let $H_i = \{x \in \hat{A}: \phi^i(x) = x\}$, $i = 1, 2, \ldots$.

Theorem 1.2.1. *For all nonempty ideals I in $C^*(\mathfrak{A}, Z)$, $I \cap \mathfrak{A} \neq \{0\}$ if and only if* interior $H_i = \emptyset$ *all i.*

Recall that the topology is that of Jacobson. An immediate consequence is the useful

Corollary. *If some nonperiodic point has a dense orbit in $\hat{A}$, then the property is true.*

Before embarking upon a proof of the theorem, we present a sequence of lemmas, at least some of which will be used elsewhere.

Definition. If ϕ is an ergodic quasi-invariant transformation on the finite measure space (X, μ), and $\mathfrak{F}$ is any second countable topology subordinate to the Borel structure, by $\operatorname{supp}_{\mathfrak{F}} \mu$ is meant the minimal closed ϕ invariant set whose complement has measure zero.

Alternatively $\operatorname{supp}_{\mathfrak{F}} \mu$ is the maximal set on which μ is diffuse, i.e. $\mu(A) > 0$ if A is nonempty and open. With this terminology, it is due to Halmos [14, p. 26] that

Lemma 1.2.1. *For almost every point in X, its orbit under ϕ is dense in* $\operatorname{supp}_{\mathfrak{F}} \mu$.

The following is also well known.

Lemma 1.2.2. *If μ is a standard measure on X, and P denotes the set of points periodic under ϕ, a quasi-invariant, ergodic Borel transformation, then $\mu(P) = 0$ or else μ consists of a finite number of atoms.*

Proof. Since μ is standard and ϕ is quasi-invariant, there exists N with $\mu(N) = 0$, and $X|N$ a standard Borel space invariant under ϕ. But then there is a second countable topology subordinate to the Borel structure on $X|N$ which is Hausdorff. Then the orbit of any periodic point is closed, and the conclusion follows for the previous lemma.

If $F \in l^1(\mathfrak{A}, Z)$, define $E_n(F) = F(n)$, $n \in Z$.

Lemma 1.2.3. *E_n is continuous in the C^*-norm on $l^1(\mathfrak{A}, Z)$ and $\|E_n\| \leq 1$.*

Proof. We recall that the C^*-norm on $l^1(\mathfrak{A}, Z)$ is $\|F\| = \sup_\rho \|\rho(F)\|$, where ρ is an irreducible representation of $l^1(\mathfrak{A}, Z)$. If $\nu \in \hat{A}$ is not periodic under ϕ, consider the transitive irreducible representation $\rho_\nu = (\pi, U_\phi)$ defined in §1. Then $H_\pi = \bigoplus_{-\infty}^{\infty} H_i$, with $H_i = H_\nu$ all i, and

$$\|\nu(F(n))\| = \sup_{\substack{\xi \in H_0; \eta \in H_n \\ \|\xi\| = \|\eta\| = 1}} |(\rho_\nu(F)\xi, \eta)|$$

so $\|\nu(F(n))\| \leq \|\rho_\nu(F)\|$.

If ν in $\hat{A}$ is periodic, of period k say, let ρ_ν^g denote the finite dimensional representation $\rho_\nu^g = (\pi, M_g U_\phi)$. If $n \equiv l \pmod{k}$ and if $\xi_i \in H_i$, $H_i = H_\nu$ all i, then

$$(\rho_\nu^g(F)\xi_0, \xi_l) = \left(\sum_{j=-\infty}^{\infty} \pi(F(j))(MgU_\phi)^j \xi_0, \xi_l\right)$$

$$= \sum_{-\infty}^{\infty} (\nu(F(l + jk))\eta_j, \xi_l) \cdot g(\nu) \cdot g(\phi(\nu)) \cdots g(\phi^l(\nu)) \cdot \lambda^j$$

where $\lambda = g(\nu) \cdots g(\phi^{k-1}(\nu))$ and $\eta_j = U_\phi^{l+jk} \xi_0$. But

$$\sup_{|\lambda| = 1} \left| \sum_{j=-\infty}^{\infty} (\nu(F(l + jk))\eta_j, \xi_l) \cdot \lambda^j \right| \geq |(\nu(F(l + jk))\eta_j, \xi_l)| \quad \text{for all } j.$$

Hence given ξ, η, there exists g for which

$$|(\rho_\nu^g(F)\xi, \eta)| \geq |(\nu(F(n))U_\phi^n \xi, \eta)|.$$

Taking sups, it has been shown that for all $\nu \in \hat{A}$, there exists ρ_ν with $\|\rho_\nu(F)\| \geq \|\nu(F(n))\|$. Hence, since $\|E(n)\| = \sup_{\nu \in A} \|\nu(F(n))\|$, the result.

Corollary ([9], [24]). *If* $F \in l^1(\mathfrak{A}, Z)$, *then* $\|F\| = 0$ *implies* $F = 0$.

Now E_n is extended by continuity to $C^*(\mathfrak{A}, Z)$. If ρ is any irreducible representation of $C^*(\mathfrak{A}, Z)$ one can attempt to define $E_n^\rho\colon \rho(C^*(\mathfrak{A}, Z)) \to \rho(\mathfrak{A})$ by $E_n^\rho(\rho(F)) = \rho(E_n(F))$. For this we need

Lemma 1.2.4. *If* $\rho = (\pi, U)$ *is an irreducible representation for which* μ_π *does not consist of a finite number of atoms, then* E_n^ρ *is both well defined and continuous, for all* $n \in Z$.

Proof. We sketch the argument which is a standard one. With the notation of the last section, it follows from Lemma 1.2.2 that $\pi = \int_X^\oplus c(x)\, d\mu_\pi$, where ϕ is a freely acting Borel isomorphism of the standard Borel space X. Using the fact that the Borel structure is both countably generated and countably separated, given any integer N and $x \in X$, one can find a Borel neighborhood W_x, with $\mu(W_x) > 0$ and $\{\phi^i(W_x)\}_{i=-N}^N$ disjoint. Thence one easily obtains $\{V_i\}_{i\in Z}$ a disjoint, measurable covering of X subordinate to $\{W_x\}_{x\in X}$. Given $\epsilon > 0$, find $H \in l^1(\mathfrak{A}, Z)$ and $N \in Z$, with $H(n) = 0$ if $|n| > N$ and $\|F - H\| < \epsilon$. Then if $\xi \in H_\pi$, and χ_S denotes the characteristic function of the set S,

$$
\begin{aligned}
\|\pi(E_0(H))\xi\|^2 &= \int_X \|c(y)(E_0(H)) \cdot \xi(y)\|^2\, dy \\
&= \sum_i \int_{V_i} \|c(y)(E_0(H)) \cdot \xi(y)\|^2\, dy \\
&\le \sum_i \|\rho(H)\|^2 \, \|M_{\chi_{V_i}} \xi\|^2 \qquad (1.3)\\
&= \|\rho(H)\|^2 \, \|\xi\|^2
\end{aligned}
$$

where (1.3) follows from

$$
\begin{aligned}
\|\rho(H) M_{\chi_{W_x}} \xi\|^2 &= \int_X \left\| \left(\sum_{-N}^{N} \pi(E_n(H)) \cdot M_{\chi_{\phi^{-n}(W_x)}} U^n \xi \right)(y) \right\|^2 d\mu(y) \\
&\ge \int_{W_x} \left\| \sum_{-N}^{N} c(y)(E_n(H)) M_{\chi_{\phi^{-n}(W_x)}}(y) \cdot (U^n\xi)(y) \right\|^2 d\mu(y).
\end{aligned}
$$

Thus $\|\pi(E_0(H))\| \le \|\rho(H)\|$. But E_0 is continuous from the previous lemma, and ρ and π are continuous, and all are of norm ≤ 1, so $\|\pi(E_0(F))\| \le \|\rho(F)\| + 2\epsilon$. Hence the lemma for $n = 0$. The general case $n = k$ is obtained by considering $\rho(H) \cdot U^{-k}$.

The converse to this last lemma is also true, as if μ_π does consist of a

finite number of atoms it is easy to define $F \in l^1(\mathfrak{A}, Z)$, for which $\rho(F) = 0$ but $\rho(E_0(F)) \neq 0$.

The following might be considered as a sort of converse to Lemma 1.2.3. It is one of the keys to understanding the structure of the covariance algebra $C^*(\mathfrak{A}, Z)$.

Lemma 1.2.5. *For $F \in C^*(\mathfrak{A}, Z)$, if $E_n(F) = 0$ for all n, then $F = 0$.*

Proof. What must be shown is that if $\rho(E_n(F)) = 0$ $\forall n, \forall \rho$, then $\rho(F) = 0$ $\forall \rho$. If $\rho = (\pi, U)$ is any irreducible representation, then clearly $\rho_\lambda = (\pi, \lambda \cdot U)$ is also, for any $\lambda \in C$, with $|\lambda| = 1$. Choose ξ, η in H_ρ. For any $H \in l^1(\mathfrak{A}, Z)$, we have

$$(\rho(H)\xi, \eta) = \sum_{i=-\infty}^{\infty} (\pi(E_i(H)) \cdot U^i \xi, \eta).$$

So we can define $h\colon S^1 \to C$, by

$$h(\lambda) = (\rho_\lambda(H)\xi, \eta) = \sum_{i=-\infty}^{\infty} (\pi(E_i(H)) \cdot U^i \xi, \eta) \cdot \lambda^i = \sum_{i=-\infty}^{\infty} a_i \lambda^i.$$

Since $\Sigma |a_i| < \infty$, $h(\lambda) \in L^2(S^1)$. If $H_k \to F$ in $C^*(\mathfrak{A}, Z)$, then from the definition of the norm in $C^*(\mathfrak{A}, Z)$ we have that

$$h_k(\lambda) \to f(\lambda) = (\rho_\lambda(F)\xi, \eta) \quad \text{uniformly in } \lambda.$$

Hence $h_k(\lambda) \to f(\lambda)$ in $L^2(S^1)$. But if $h_k(\lambda) = \Sigma_{i=-\infty}^{\infty} a_{ik}\lambda^i$, and $f(\lambda) = \Sigma f_i \lambda^i$ then $a_{ik} \to f_i$. However $a_{ik} = (\pi(E_i(H_k))\xi, \eta)$, which by Lemma 1.2.3 converges to zero, each i. Hence $f_i = 0$ for all i, so $f(\lambda) = (\rho_\lambda(F)\xi, \eta) \equiv 0$. In particular $(\rho(F)\xi, \eta) = 0$. Since ξ, η and ρ are all arbitrary, we conclude that $F = 0$.

It is an immediate consequence of the last two lemmas that

Corollary. *If $\rho = (\pi, U)$ is an irreducible representation of $C^*(\mathfrak{A}, Z)$, then ρ is faithful if and only if μ_π is not periodic and π is faithful.*

Remark. From here, when we speak of the support of μ_π, we shall mean with respect to the Jacobson topology. That we may do so and that $\operatorname{supp} \mu_\pi = \{\pi_0\colon \ker \lambda_0 \supset \ker \pi\}$ is shown in [13]. Thus π is faithful if and only if $\operatorname{supp} \mu_\pi = A$.

We turn now to the proof of the theorem. Let $H_i = \{a \in \hat{A}\colon \phi^i(a) = a\}$. Let $P = \bigcup_{i=1}^{\infty} H_i$. Let I be a selfadjoint ideal in $C^*(\mathfrak{A}, Z)$. We want to show that if interior $H_i = \emptyset$, all i, then $I \cap \mathfrak{A} \neq \{0\}$.

Any ideal I is uniquely determined as the kernel of a certain family of irreducible representations $\{\rho_\gamma\}_{\gamma \in S}$ of $C^*(\mathfrak{A}, Z)$. Let $\{\mu_\gamma\}_{\gamma \in S}$ be the cor-

responding measures on A. If $\operatorname{supp} I = \overline{\bigcup_{\gamma \in S} \operatorname{supp} \mu_\gamma} \neq \widehat{A}$, then from the definition of the topology, for some $T \neq 0$ in $\mathfrak{A}$, $\pi(T) \equiv 0$ for all π in $\operatorname{supp} I$. Thus $T \in I$.

So assume $\operatorname{supp} I = \widehat{A}$. Let $F \neq 0$ be in I. For all nonperiodic ρ_γ, $\rho_\gamma(F) = 0$ implies $\rho_\gamma(E_n(F)) = 0$ for all n by Lemma 1.2.4. Let $\mathfrak{N} = \{\gamma: \rho_\gamma$: is not periodic$\}$, and $Y = \overline{\bigcup_{\gamma \in \mathfrak{N}} \operatorname{supp} \mu_\gamma}$. From the above it follows that $\pi(E_n(F)) = 0$ for all π in Y and all n in Z. Thus if $F \neq 0$, it follows from Lemma 1.2.5 that for some π_0 in Y^c and integer n_0, $\pi_0(E_{n_0}(F)) \neq 0$. There are now two cases to be considered.

Suppose first that π_0 is periodic. Using the fact that I is an ideal it can be assumed that $\|\pi_0(E_0(F))\| = 2$ say. Since $\operatorname{supp} I = A$, it must be that $Y^c \subset \overline{P}$, and the hypothesis that the interior of H_i is empty for all i implies that either

(i) π_0 is the limit of a net $\{\pi_d\}_{d \in D}$ of nonperiodic points,

or that

(ii) it is the limit of a sequence of periodic points whose periods tend to infinity.

This makes use of the fact that ϕ is a homeomorphism in the Jacobson topology. Since Y^c is open and contained in $\overline{P}$, if (i) holds then $\{\pi_d\}_{d \in D}$ is ultimately in $\overline{P}$. But each π_d is not periodic, a net $\{\nu_l\} \subset P$ can be chosen with $\nu_l \to \pi_0$ and period $\nu_l \to \infty$. This is (ii).

Since $Y^c \subset \overline{\bigcup_{\gamma \in M \setminus \mathfrak{N}} \operatorname{supp} \mu_\gamma}$, if (ii) holds then the net $\{\pi_d\}$ can in fact be chosen so that the representations $\rho_d = (\pi_d, M_{g_d} U_\phi)$ are each in $\{\rho_\gamma\}_{\gamma \in M \setminus \mathfrak{N}}$ for some g_d.

On the other hand, if π_0 is not periodic then since $\pi_0 \in \overline{P}$ implies that for some collection $\{\pi_\alpha\} \subset P$, $\ker \pi \supset \bigcap \ker \pi_\alpha$, it follows that $\pi_0(E_0(F)) \neq 0$ implies $\pi_{\alpha_0}(E_0(F)) \neq 0$, some α_0, and we are in the first case above.

Choose $A \in l^1(\mathfrak{A}, Z)$ with $\|A - F\| < \frac{1}{4}$, and choose $N_0 \geq 0$ with $\|A\|_1 \leq \sum_{-N}^{N} \|E_n(A)\| + \frac{1}{4}$. If $\rho = (\pi, M_f U_\phi)$ is a transitive irreducible representation of $C^*(\mathfrak{A}, Z)$, based on the orbit of π_0, where π_0 has period k, and if $\lambda = f(\pi_0) \cdot f(\phi(\pi_0)) \cdots f(\phi^{k-1}(\pi_0))$, then for any $\xi, \eta \in H_{\pi_0}$ with $\|\xi\| = \|\eta\| = 1$,

$$\left|\left(\sum_{i=-\infty}^{\infty} \pi_0(E_{0+ik}(A)) \cdot \lambda^i, \xi, \eta\right)\right| = |(\rho(A)\xi, \eta)| \leq \|\rho(A)\|.$$

Thus

$$|(\pi_0(E_0(A))\xi, \eta)| \leq \|\rho(A)\| + \left|\sum_{i \neq 0} (\pi_0(E_{ik}(A))\xi, \eta)\right|.$$

So

$$\|\pi_0(E_0(A))\| \le \|\rho(A)\| + \sum_{i\neq 0} \|\pi_0(E_{ik}(A))\|$$
$$\le \|\rho(A)\| + \tfrac{1}{4} \quad \text{if } k > N_0.$$

Thus ultimately

$$\|\pi_d(E_0(A))\| \le \|\rho(A)\| + \tfrac{1}{4} \le 2$$

since

$$\|\rho_d(A)\| = \|\rho_d(A) - \rho_d(F)\| \le \|A - F\| = \tfrac{1}{4}.$$

But then by upper continuity [8, Proposition 3.3.2], it follows that $\|\pi_0(E_0(A))\| \le \tfrac{1}{2}$, thence $\|\pi_0(E_0(F))\| \le \tfrac{3}{4}$, a contradiction. This is the desired result since it has been shown that if interior $H_i = \emptyset$ all i, then supp $I = \hat{A}$ implies that $I = \{0\}$.

For the converse, suppose interior $H_{j_0} \neq \emptyset$, then some π_0 has an open neighborhood $W_{\pi_0} \subset H_{j_0}$. This means that for some $A \in \mathfrak{A}$, $\pi(A) = 0$ for all π in the complement of W_{π_0}, but $\pi_0(A) \neq 0$. For each a in $\hat{A}$, let $\rho_a = (\pi_a, U_\phi)$ denote the transitive irreducible representation based on its orbit. Since $\pi_a(A) = 0$ all a implies $A = 0$, it suffices to find $F \neq 0$ in $l^1(\mathfrak{A}, Z)$, with $\rho_a(F) = 0$ all a in $\hat{A}$.

If a has period k and $F \in l^1(\mathfrak{A}, Z)$, then certainly $\rho_a(F) = 0$ if $\Sigma_{n=-\infty}^{\infty} \phi^l(a)(F(i + nk)) = 0$, $0 \le i, l \le k - 1$. Define

$$F(n) = 0 \qquad \text{if } |n| > j_0^2 + 1,$$
$$F(n) = v_n A \quad \text{if } |n| < j_0^2 + 1,\ v_n \in R.$$

Then since every point in W_{π_0} has period at most j_0, it follows that any nontrivial solution for the v_n's to a system of at most $j_0(j_0 + 1)/2$ equations gives an F with the desired properties.

We may notice that Lemma 1.2.4 implies that every irreducible representation of a covariance algebra $C^*(\mathfrak{A}, Z)$ is in fact a covariance algebra; $C^*(\pi(\mathfrak{A}), Z_k)$ if π is periodic and $C^*(\pi(\mathfrak{A}), Z)$ otherwise. We also have

Theorem 1.2.2. *Let $\rho = (\pi, U)$ be any irreducible representation of $C^*(\mathfrak{A}, Z)$. If I is any nonzero selfadjoint ideal in $\rho(C^*(\mathfrak{A}, Z))$, then $I \cap \pi(\mathfrak{A})$ $\neq \{0\}$.*

Proof. If π is periodic, then $\rho(C^*(\mathfrak{A}, Z)) \approx \pi(\mathfrak{A}) \otimes M_k$, where $M_k = k \times k$ matrices, and the result is clear. If π is not periodic, then $\mu_\pi(P) = 0$, by Lemma 1.2.2, and μ_π is diffuse on supp μ_π, so interior $H_i' = \emptyset$ each i, where $H_i' = H_i \cap \operatorname{supp} \mu_\pi$. Thus the theorem applies.

1.3. Type I covariance algebras. For a separable C^*-algebra $\mathfrak{A}$, the following are equivalent [2], [8]:

(i) Every irreducible representation of $\mathfrak{A}$ contains the compact operators.

(ii) Any two irreducible representations of $\mathfrak{A}$ with the same kernel are unitarily equivalent.

(iii) Every factor representation of $\mathfrak{A}$ is of type I, i.e. $\pi(\mathfrak{A})''$ is a type I von Neumann algebra.

If any of these hold, $\mathfrak{A}$ is said to be G.C.R. or type I. Takesaki [23] and Zeller-Meier [24] have given conditions under which a general covariance algebra is G.C.R. Utilizing the idea of "induced representations", it is shown that $C^*(\mathfrak{A}, Z)$ is G.C.R. if and only if $\mathfrak{A}$ is G.C.R. and the action of ϕ on $\hat{A}$ is smooth or regular in the sense of Mackey [20]. Glimm [12] and Effros [11] give lists of conditions under which this is true. Since the group here is Z this result can be stated more simply than in general.

Definition. A point $a \in \hat{A}$ is said to be discrete in its orbit (under ϕ) if $\phi^{n_i}(a) \to a$ as $n_i \to +\infty$ or $-\infty$ implies a is periodic (with respect to ϕ).

Theorem 1.3.1. *If $\mathfrak{A}$ is G.C.R., then the following are equivalent:*

(i) *$C^*(\mathfrak{A}, Z)$ is G.C.R.*

(ii) *Every a in $\hat{A}$ is discrete in its orbit.*

(iii) *No two orbits have the same closure.*

(iv) *Every quasi-invariant ergodic measure on $\hat{A}$ is transitive.*

Proof. (i) $\Rightarrow$ (ii) If $C^*(\mathfrak{A}, Z)$ is G.C.R., if $a \in \hat{A}$ is not periodic, and if ρ_a is the transitive irreducible representation based on the orbit of a, then $\rho_a(C^*(\mathfrak{A}, Z)) \supset C(H_{\rho_a})$, so by Theorem 1.2.2, for some $A \in \mathfrak{A}$, $\rho_a(A)$ is a rank one projection. In general $\pi_1(A) \oplus \pi_2(A)$ can never be rank one unless $\pi_1(A)$ or $\pi_2(A)$ is zero. We conclude that $\rho_a(A)$ can be rank one only if $\phi^i(a)(A) = 0$ for all $i \neq i_0$. But if $\phi^{n_i}(a) \to a$ this is impossible. Hence (ii).

(ii) $\Rightarrow$ (iii) If (ii) holds and y_1 and y_2 contradict (iii), then $\phi^{n_i}(y_1) \to y_1$ some $\{n_i\}$ and $\phi^{m_i}(y_2) \to y_1$ some $\{m_i\}$. Then that $\phi^{l_i}(y_1) \to y_1$ some $\{l_i\}$, l_i distinct, is immediate.

(iii) $\Rightarrow$ (i) If (iii) holds then no two transitive representations have the same kernel unless they are unitarily equivalent. So if (i) does not hold then there exists an intransitive measure μ. But then it follows from Lemma 1.2.1 that there must be at least two distinct orbits which have supp μ as their closure.

(i) $\Leftrightarrow$ (iv) In the course of the last part, it was shown that (i) $\Rightarrow$ (iv). But from the description given in §1 of the factor representations of $C^*(\mathfrak{A}, Z)$ it follows that if every quasi-invariant ergodic measure is transitive, then every factor representation is the direct sum of irreducibles and hence of type I. Thus (iv) $\Rightarrow$ (i).

1.4. A representation theorem for certain operators with closed range. If $A \in L(H)$, as usual H separable, then among the many ways of expressing the fact that the range of A is closed are the following [7]:

(i) $\exists C > 0$ such that $\|A^*Ax\| \geq C\|x\|$, $\forall x \in (\ker A)^{\perp}$.

(ii) The origin is an isolated point of the spectrum of A^*A.

This last shows in particular that if A has closed range, then every representation of A does. If $A = UD$ is the polar decomposition [2], with U a partial isometry and D a positive operator, then by using (ii) to define a contour integral it is shown in [6] and [7] respectively that U^*U and U belong to $C^*(A)$. We shall also need the following. The proof is entirely straightforward so we omit it.

Lemma 1.4.1. *If π is a representation of a C^*-algebra $\mathfrak{A} \subset L(H)$, and if $A \in \mathfrak{A}$ has polar decomposition $A = UD$ and closed range, then $\pi(A)$ has polar decomposition $\pi(A) = \pi(U) \cdot \pi(D)$.*

Suppose $A = UD$ has closed range. Let ϕ and ϕ^{-1} denote the continuous linear maps of $C^*(A)$ into itself, defined for B in $C^*(A)$ by $\phi(B) = UBU^*$ and $\phi^{-1}(B) = U^*BU$.

Let $\mathfrak{D} = C^*(\{\phi^n(D), U^{(-n)}U^{(n)}\}_{n \in Z}) \subset C^*(A)$, where the notation is $U^{(n)} = U^n$ if $n \geq 0$ and $U^{(n)} = (U^*)^{-n}$ if $n \leq 0$. An alternative description of $\mathfrak{D}$ is that it is the minimal C^*-algebra containing D and 1 and invariant under both ϕ and ϕ^{-1}. Let I_U denote the closed selfadjoint ideal in $C^*(A)$ generated by $U^*U - UU^*$, and let q denote the canonical quotient map from $C^*(A)$ into $C^*(A)/I_U$. Then

Theorem 1.4.1. *If (C) holds, namely if for every finite collection $\{D_n\}_{n \in J} \subset \mathfrak{D}$, $\Sigma_{n \in J} D_n U^{(n)} \in I_U$ implies that each $D_n \in I_U$, then $q(C^*(A))$ is $*$-isomorphic to a covariance algebra $C^*(q(\mathfrak{D}), Z)$.*

Proof. Since $q(C^*(A))$ is a C^*-algebra, we can find a $*$-isomorphism π of $q(C^*(A))$ into some $L(H_\pi)$, and it may be assumed that π is nondegenerate [8]. For brevity, let B' denote the image of any $B \in C^*(A)$ under the representation $\pi \circ q$. Then by Lemma 1.4.1, we have that $A' = U'D'$ is the polar decomposition. Since $q(U)$ is normal, U' is also, so $N(A') = N(U') = N(U'^*) = N(A'^*)$, and the assumption of nondegeneracy implies that U' is unitary. If $\mathfrak{D}' = \pi \circ q(\mathfrak{D})$, then $\mathfrak{D}'$ is the minimal C^*-algebra containing 1 and D', and invariant under ϕ': $B' \to U'B'U'^*$ and ϕ^{-1}: $B' \to U'^*B'U'$. Thus $q(C^*(A))$ is $*$-isomorphic to $C^*(A')$, $q(\mathfrak{D})$ to $\mathfrak{D}'$, and ϕ' and ϕ'^{-1} are inverse $*$-automorphisms of $\mathfrak{D}'$ implemented by the unitary operator U'. Additionally

it follows (C$'$) that $\Sigma_{n\in J}D'_n U'^n = 0$ implies $D'_n = 0$ each n in J. The primes will now be omitted.

Let B denote the *-subalgebra of $C^*(A)$ consisting of all elements of the form $\Sigma_{n\in J} D_n U^n$, for some finite subset J of Z, and each D_n belonging to $\mathfrak{D}$. Since $A = UD = (UDU^{-1}) \cdot U \in B$, B is clearly dense in $C^*(A)$. Let the action of Z on $\mathfrak{D}$ be given by the automorphism ϕ, and define χ: $B \to l^1(\mathfrak{D}, Z)$ by $\chi(\Sigma_{n\in J} D_n U^n)(k) = \delta_{k,J} D_k$ where $\delta_{k,J}$ is the Kronecker delta. This map is well defined since (C$'$) guarantees the uniqueness of the representation of elements of B. It is clearly linear and into $l^1(\mathfrak{D}, Z)$. The verifications that χ is multiplicative, *-preserving, and 1-1 are all routine. Defining χ^{-1} on the range of χ we see that it is continuous in the l^1 norm, so that χ^{-1} can be extended to a representation of $l^1(\mathfrak{D}, Z)$. But since this representation is faithful on $\mathfrak{D}$, the corollary to Lemma 1.2.5 gives the conclusion.

PART II. BILATERAL WEIGHT SHIFTS

2.1. Uniqueness of the basis. Let H be a separable Hilbert space. Then V in $L(H)$ is a bilateral weighted shift means that for some orthonormal basis $\{e_n\}_{n\in Z}$, $Ve_n = d_n e_n$, where d_n are complex scalars. It is well known that such an operator is reducible if and only if some $d_i = 0$ or $\{d_i\}$ is a periodic sequence, and that two such are unitarily equivalent if $|d_i| = |d'_i|$ each i [15], [18]. Thus we restrict ourselves to the case $d_i > 0$.

Theorem 2.1.1. *The basis $\{e_n\}$ with respect to which V has this shift form is unique if and only if V is irreducible.*

Proof. First if V is the ordinary unweighted bilateral shift, i.e. multiplication by z on $L^2(S^1)$, and if $\phi(z)$ is any inner function then $\{z^n\phi\}_{n\in Z}$ form an alternative basis. If V is periodic of period k one simply has to consider $\{z^n\phi(z^k)\}_{n\in Z}$. The converse can be shown by a direct argument or it can be considered a particular case of a more general situation. For this let (X, μ, ϕ), and (Y, ν, χ) be triples of (standard Borel space, finite measure, quasi-invariant Borel isomorphism). For $g > 0$ and $f > 0$ in $L^\infty(Y, \nu)$ and $L^\infty(X, \mu)$ respectively, let $M_g U_\chi$ and $M_f U_\phi$ denote the obvious weighted translation operators on $H_1 = L^2(Y, \nu)$ and $H_2 = L^2(X, \mu)$, respectively.

Lemma 2.1.1. *If $M_f U_\phi$ and $M_g U_\chi$ are irreducible and are unitarily equivalent via $V \in L(H_1, H_2)$ then there exists a nonsingular Borel isomorphism λ of (X, μ) onto (Y, ν) such that $\lambda \circ \phi = \chi \circ \lambda$ and $V = U_\lambda$.*

The theorem now follows if two distinct bases are considered as $L^2(Z, \mu_1)$ and $L^2(Z, \mu_2)$.

For the proof of the lemma, let $\mathfrak{D}_1 = C^*(\{M_{f\circ\phi^n}\}_{n\epsilon Z})$ and $\mathfrak{D}_2 = C^*(\{M_{g\circ\chi^n}\}_{n\epsilon Z})$. Let $\mathfrak{A}_1 = C^*(M_g U_\phi)$, and $\mathfrak{A}_2 = C^*(M_g U_\chi)$. Since the constant function $h \equiv 1$ must be cyclic for the irreducible $M_f U_\phi$, it follows that the C^*-algebra of functions in $\mathfrak{D}$ must have all of $L^2(X, \mu)$ as its L^2 closure. Since μ is finite and we have a C^*-algebra of functions in L^∞, one can readily conclude that every function h in L^2 can be approximated almost everywhere by a sequence h_n of L^∞ functions bounded in L^∞ norm. Then one obtains that $L_{h_n} \to L$ strongly. Thus the von Neumann algebra generated by $\mathfrak{D}_1$ is in fact all of $L^\infty(X, \mu)$. Similarly for $\mathfrak{D}_2$ and $L^\infty(Y, \nu)$. Since $M_f U_\phi$ and $M_g U_\chi$ are the respective polar decompositions, one has $V\mathfrak{D}_1 V^{-1} = \mathfrak{D}_2$, and thence by the above remarks $VL^\infty(X, \mu)V^{-1} = L^\infty(Y, \nu)$. Since we have standard measure algebras, the results of [16] imply that the isomorphism is implemented by a nonsingular point transformation λ. The remaining conclusions of the lemma follow easily.

We should like to thank L. G. Brown for discussions related to the above.

2.2. The generated C^*-algebra and the canonical diagonal spectrum. Let the bilateral weighted shift V have polar decomposition $V = UD$, where D is the diagonal operator $De_n = d_n e_n$, and U is the bilateral shift. Using the terminology of Theorem 1.4.1, $I_U = \{0\}$, and $\phi^n(D) = UDU^{-n}$, so $\mathfrak{D} = C^*(\{\phi^n(D)\}_{n\epsilon Z})$ consists of diagonal operators. Let X denote the maximal ideal space or spectrum of the commutative C^*-algebra $\mathfrak{D}$. Let ϕ also denote the induced homeomorphism of X. Then since condition (C) clearly holds, we have

Theorem 2.2.1. *If V is a bilateral weighted shift with closed range, then $C^*(V)$ is $*$-isomorphic to the covariance algebra $C^*(X, \phi)$.*

If $\{V_\gamma = UD_\gamma\}_{\gamma\epsilon\mu}$ is a collection of bilateral weighted shifts, we may assume that at least one V_{γ_0} has positive weights, so that $V_{\gamma_0} = UD_{\gamma_0}$ with U the bilateral shift and D diagonal is the polar decomposition. Let $\phi(A) = UAU^{-1}$ for $A \in L(H)$ and $\mathfrak{D} = C^*(\{\phi^n(D_\gamma)\}_{n\epsilon Z, \gamma\epsilon\mu})$. Denoting the spectrum of $\mathfrak{D}$ by X and the induced homeomorphism of X by ϕ also, in a manner entirely analogous to the last theorem, one obtains

Theorem 2.2.2. *If V_{γ_0} has closed range then $C^*(\{V_\gamma\}_{\gamma\epsilon\mu})$ is $*$-isomorphic to $C^*(X, \phi)$.*

Returning to the case of one shift, a natural question is which pairs (X, ϕ) can arise under the correspondence in Theorem 2.2.1. If $n \in Z$, define w_n in $\mathrm{Hom}(\mathfrak{D}, C) = X$ by $w_n(A) = a_n$ if $A = \mathrm{diag}\{a_i\}_{i\epsilon Z}$. Then $\phi^j(w_n)$

$= w_{n-j}$, and $\{w_n\}_{n \in Z}$ is dense in X, since if $f \in C(X)$, with $f(w_n) = 0$, all n, then the inverse Gelfand transform of f is zero, and hence $f \equiv 0$. So the pair (X, ϕ) has the property that there exists an orbit of ϕ which is dense in X.

Let Y denote the spectrum of D. Define $T: X \to \Pi_{-\infty}^{+\infty} Y$, by

$$T(x) = (\dots, x(\phi(D)), x(D), x(\phi^{-1}(D)), \dots).$$

If $T(X)$ is given the topology induced by the product topology on ΠY, then since $\mathfrak{D}$ is generated as a C^*-algebra by $\{\phi^n(D)\}_{n \in Z}$, T is both continuous and 1-1. Consequently since X is compact, it is a homeomorphism of X onto $T(X)$. $T(X)$ will be denoted by X_c and referred to as the *canonical form* of the *diagonal spectrum* X. Note that under T, ϕ becomes the usual right shift on a product space.

Theorem 2.2.3. *If ϕ is a homeomorphism of a compact Hausdorff space X, then there exists a bilateral weighted shift V, with $C^*(V)$ naturally *-isomorphic to $C^*(X, \phi)$ if and only if*

(i) *there exists a point $x_0 \in X$, with dense orbit under ϕ, and*

(ii) *there exists f in $C(X, R)$, such that $\{f \circ \phi^n\}_{n \in Z}$ separates points in X.*

Proof. The necessity of (i) has been shown. For (ii), let $p_0 \in C(X_c, R)$ be the projection on the zero coordinate and pull back to X. Conversely, assume first that x_0 is not periodic under ϕ. If $f \in C(X, R)$ satisfies (ii), let $f'(x) = f(x) + 2\|f\|$. Let μ be a transitive ergodic measure on the orbit of x_0, and let $\rho_{x_0} = (\pi_\mu, U_\phi)$ be the corresponding representation of $C^*(X, \phi)$. By the corollary to Lemma 1.2.5, ρ is a faithful representation and if V is a bilateral weighted shift with weights $d_n = f'(\phi^n(x_0))$, then V is the image under ρ of an obvious $F \in C^*(X, \phi)$ and has positive weights and closed range, so an application of the Stone-Weierstrass theorem gives the desired conclusion. If x_0 is periodic, X consists of a finite number of points, say k, and any shift with nonzero weights of period k will clearly suffice.

From the proof it is immediate that

Corollary 1. *$C^*(X, \phi)$ is naturally *-isomorphic to a C^*-algebra generated by a family of bilateral weighted shifts if and only if ϕ has a dense orbit.*

We also have, using the real and imaginary parts as in the theorem,

Corollary 2. *If ϕ has a dense orbit, then $C^*(X, \phi)$ is *-isomorphic to a*

C^*-algebra generated by two bilateral weighted shifts if there exists f in $C(X, C)$ such that $\{f \circ \phi^n\}_{n \in Z}$ separates points in X.

This enables us to give an example of a C^*-algebra generated by a pair of shifts, which is not generated by a single one.

Example 1. With $S^1 = \{z: z \in C, |z| = 1\}$. Let $X = \Pi_{-\infty}^{\infty} S^1$ and ϕ be the usual shift. This pair clearly satisfies the conditions of Corollary 2, but if for $f \in C(X, R)$, $\{f \circ \phi^n\}_{n \in Z}$ separates points, then letting T: $X \to \Pi_{-\infty}^{\infty} f(X)$ be given by $T(x) = (\ldots, f(\phi(x)), f(x), f(\phi^{-n}(x)), \ldots)$ then T is a homeomorphism onto its range. Now $\chi = T \circ \phi \circ T^{-1}$ is the usual shift on $T(X) \subset \Pi_{-\infty}^{\infty} f(X)$, and the set of fixed points of ϕ must be homeomorphic to those of Y. But this gives T homeomorphic to $f(X) \subset R$, which is impossible.

Some examples of how the theorem itself applies follow:

Example 2. Let Y be any compact subset of the real line, and $X = \Pi_{-\infty}^{\infty} Y$ with the product topology. Let ϕ be the usual shift. Since, with respect to the natural product measure, ϕ is measure preserving and ergodic, by the lemma of Halmos quoted earlier, Lemma 1.2.1, almost every point has dense orbit, and clearly any coordinate function works.

Example 3. Let X be the n-torus, and let ϕ be an ergodic rotation of this topological group. Again from Lemma 1.2.1, this time with respect to Haar measure, almost every point has dense orbit. In fact, every point must have. Define $f \in C(X, R)$ by $f(z_1, z_2, \ldots, z_m) = \mathrm{Re}(z_1 + z_2 + \ldots + z_m)$. A simple argument using the fact that a Vandermonde matrix is invertible if and only if the elements are distinct [17] shows that $\{f \circ \phi^n\}_{n \in Z}$ separates points.

Returning to the canonical diagonal spectrum, recall that we had T: $X \to X_c$ by (2.2). So for the dense subset $\{w_n\}_{n \in Z}$ in X, that was defined previously, we have $T(w_k) = (\ldots, d_{k-1}, d_k, d_{k+1}, \ldots)$. (Recall $V = UD$, with $D = \mathrm{diag}\{d_i\}_{i=2}$.) Thus X_c is simply the closed subset of $\Pi_{-\infty}^{\infty}(\text{spectr. } D)$ generated by D and its translates. And if χ is the usual shift, V is algebraically equivalent to the element F_V in $C^*(X_c, \chi)$ of the form

$$\begin{aligned} F_V(n) &= p_0 \quad \text{if } n = 1, \\ &= 0 \quad \text{if } n \neq 1. \end{aligned}$$

Thus at least when V is G.C.R. it is a simple matter using the representation theory described earlier to write down the irreducible representations of V.

Example. Let $V = UD$ with

$$D = \mathrm{diag}\{\ldots, 1, 1, \dot{1}, 1, 2, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, \ldots\}.$$

Then X_c is easily determined and one ascertains that V has, to within uni-

tary equivalence, only the following distinct irreducible representations:

(i) the identity representation,

(ii) a representation as a shift with weights $\{\ldots, 1, 1, 1, 2, 2, 2, \ldots\}$,

(iii) as a shift with weights $\{\ldots, 2, 2, 2, 1, 1, 1, \ldots\}$, and

(iv) for each $\lambda \in \mathbf{C}$, a one dimensional representation as λ and 2λ.

2.3. **Shifts without closed range.** It was necessary that V have closed range in Theorem 2.2.1 in order that the diagonal algebra $\mathfrak{D}$ be a subalgebra of $C^*(V)$ and so that for all $B \in \mathfrak{D}$, we should have $B \cdot U^k$ in $C^*(V)$ for all $k \in Z$. In general if V does not have closed range, one can only say that $C^*(V)$ is *-isomorphic to some subalgebra of $C^*(X, \phi)$. For certain shifts, one can say more. If V is essentially (or almost) normal i.e. $V^*V - VV^* \in C(H)$, then except for $D = \lambda 1$, which we ignore, V is irreducible, so $C^*(V) \supset C(H)$ [2]. If $K(H)$ denotes the compact diagonal operators, then $\mathfrak{D} = C^*(D) + K(H) \subset C^*(V)$, and clearly $\mathfrak{D}^1 \cdot U^k \subset C^*(V)$ for $\mathfrak{D}^1 = \mathfrak{D} - \{\lambda I\}_{\lambda \in C}$, so the proof of Theorem 1.4.1 shows that $C^*(V)$ is *-isomorphic to $C^*(X_0, \phi)$, with X_0 only locally compact here. In particular, every representation ρ consists of a pair (π, L), and $\rho(V) = \pi(D) \cdot L$. For an arbitrary shift this is no longer true. In fact, one can show quite easily

Theorem 2.3.1. *$C^*(V = DU)$ has the property that for every irreducible representation ρ, $\rho(V) = \pi(D) \cdot L$ for some representation π of $\mathfrak{D}$ and unitary operator L if and only if*

$$\text{(O)} \qquad d_{n_i} \longrightarrow 0 \ \textit{implies}\ d_{n_{i+k}} \longrightarrow 0, \quad \textit{all } k \in Z.$$

One can go further and show that

Theorem 2.3.2. *$C^*(V)$ is a covariance algebra $C^*(X, \phi)$ if and only if V satisfies condition* (O).

Proof. The previous theorem shows the necessity. Let $\mathfrak{D}^\#$ be the subalgebra of $C(V)$ consisting of all diagonal operators. Let $A_n = (V^n V^{*n})^{1/2}$ and $B_n = (V^{*n}V^n)^{1/2}$, all $n \geq 1$. As always $\mathfrak{D} = C^*(\{\phi^n(D)\}_{n \in Z}) \approx C(X)$. If $w \in X$, then it follows from (O) that $w(\phi^n(D)) \neq 0$, all n. But then using the fact that A_n and B_n are in $\mathfrak{D}^\#$, the "functions" in $\mathfrak{D}^\#$ are seen to separate the points of X. $\mathfrak{D}^\#$ is a *-algebra, so the Stone-Weierstrass theorem gives $\mathfrak{D}^\# = \mathfrak{D}$. If $\mathfrak{D}^\#_n$ denotes $\{B \in \mathfrak{D}\colon B \cdot U^n \in C^*(V)\}$, then $\mathfrak{D}^\#_n$ contains the selfadjoint algebra $D \cdot \mathfrak{D}$, the elements of which also must separate the points of X. Thus $\mathfrak{D}^\#_n = \mathfrak{D}$, each n, and the conclusion of the theorem follows.

2.4. **Algebraic equivalence.** We turn now to the question of when two irreducible bilateral weighted shifts V_1 and V_2 are algebraically equivalent,

i.e. when does there exist a faithful representation π of $C^*(V_1)$ with $\pi(V_1) = V_2$. The first lemma shows that we may assume that V_1, and hence necessarily V_2, have closed range.

Lemma 2.4.1. *Let $A \in L(H)$ have polar decomposition $A = UD$. If π is any representation of $C^*(A)$ for which $N(\pi(A)) = N(\pi(A^*)) = \{0\}$, then π has an extension to a representation of $C^*(A, U)$ on the same Hilbert space H_π.*

Proof. From [8, Proposition 2.10.2] there exists an extension π' of π to $C^*(A, U)$ such that $H_{\pi'} \supset H_\pi$ and $\pi'(B)|_{H_\pi} = \pi(B)$, for all B in $C^*(A)$. Then $\pi'(A)|_{H_\pi} = \pi'(U) \circ \pi'(D)|_{H_\pi}$ implies, since $\overline{\pi'(D)H_\pi} = H_\pi$ by hypothesis, that $\pi'(U)H_\pi \subset H_\pi$. Similarly considering $\pi'(A^*)$, one obtains $\pi'(U^*)H_\pi \subset H_\pi$. So π' is reduced by H_π and the corresponding subrepresentation is the desired one.

Returning to the irreducible weighted shifts V_1 and V_2, we may now assume closed range. Let X_{ic}, $i = 1, 2$, denote their respective canonical diagonal spectrum. Then

Theorem 2.4.1. *V_1 is algebraically equivalent to V_2 if and only if $X_{1c} = X_{2c}$.*

Proof. By Theorem 2.2.1, each $C^*(V_i)$ is naturally *-isomorphic to $C^*(X_{ic}, \chi)$ where χ is the usual shift on a product space, and under this *-isomorphism, V_i is carried to F_i in $C^*(X_{ic}, \chi)$, where $F_i(1) = p_0$, the zero coordinate projection and $F_i(n) = 0$ if $n \neq 1$. Hence the sufficiency.

If V_1 is algebraically equivalent to V_2, then $C^*(V_2)$ is the image of a faithful irreducible representation ρ of $C^*(X_{ic}, \chi)$ with $\rho(F_1) = V_2$. By §1.1, if $\rho = (\pi, L)$, then μ_π is transitive, and if it is based on the orbit of x_0, then $\rho(F_1)$ is a shift whose weights have as their absolute values the coordinates of x_0. But this sequence and its translates generates X_{2c}, so $X_{2c} \subset X_{1c}$, and symmetry reverses the inequality.

As previously remarked, if $X = \Pi_{-\infty}^{\infty}\{1, 2\}$, and χ is the usual shift, then with respect to the usual product measure, almost every point has dense orbit. In this sense

Corollary. *Almost all bilateral weighted shifts whose weights are* 1 *or* 2 *are algebraically equivalent.*

Remark. If the weights of a particular shift form an almost periodic function on Z [19], then the diagonal spectrum X is a topological group. There is a natural homeomorphism of X onto a subgroup of $\Pi_{-\infty}^{\infty} S^1$ and it is possible

to formulate a condition for algebraic equivalence in terms of this subgroup.

2.5. N.G.C.R. shifts. Recall that a C^*-algebra is said to be N.G.C.R. if it has no C.C.R. ideal [2]. So if V is any irreducible operator, $C^*(V)$ is N.G.C.R. means simply that $C^*(V) \cap C(H) = \{0\}$. If V has closed range then $C^*(V) \approx C^*(X, \phi)$ and known conditions apply [24]. In fact, these conditions are derived in Part I. Actually the same condition applies whether or not the range is closed. Let $V = UD$, where $D = \operatorname{diag}\{d_k\}_{k \in Z}$.

Theorem 2.5.1. *$C^*(V) \cap C(H) = \{0\}$ if and only if there exists $n_i \to \infty$ with $d_{n_i+k} \to d_k$, all $k \in Z$.*

Proof. We know that $C^*(V)$ is *-isomorphic to some subalgebra of $C^*(X, \phi)$, where X is the spectrum of $\mathfrak{D} = C^*(\{\phi^n(D)\}_{n \in Z})$ and $\phi(D) = UDU^*$. Let $\mathfrak{D}^\#$ be the C^*-subalgebra of $C^*(V)$ consisting of all diagonal operators. Let $X^\#$ denote its spectrum. We shall see that $\mathfrak{D}^\#$ is a "large enough" subalgebra of $\mathfrak{D}$, so that we can work with it.

Clearly, if the weights are periodic, $C^*(V) \cap C(H) = \{0\}$, so we may restrict ourselves to the case of V irreducible. Then $C^*(V) \cap C(H) \neq \{0\}$ implies $C^*(V) \cap C(H) \equiv C(H)$ [2]. Thus $\mathfrak{D}^\#$ is a representation π of $C(X^\#)$ and so from the well-known structure of these [2], $\mathfrak{D}^\#$ will contain rank one operators if and only if

(i) π is multiplicity free, and

(ii) μ_π has an isolated atom.

For all $k \in Z$, define w_k' in $X^\#$ by $w_k'(B) = b_n$ for $B = \operatorname{diag}\{b_n\}_{n \in Z}$. Then $\{w_k'\}_{k \in Z}$ is dense in $X^\#$, and it is seen that the measure μ_π is purely atomic with atoms $\{w_k'\}_{k \in Z}$.

Put $A_n = \sqrt{V^n V^{*n}} = \phi(D) \cdot \phi^2(D) \cdots \phi^n(D)$, and $B_n = \sqrt{V^{*n}V^n} = D \cdot \phi^{-1}(D), \ldots, \phi^{-n+1}(D)$. Since these all belong to $\mathfrak{D}^\#$, by the continuing assumption of nonzero weights each w_k' has a *unique* extension to the obvious w_k in X. In particular $w_i' = w_j' \Rightarrow w_i = w_j \Rightarrow w^i(\phi^k(D)) = w^j(\phi^k(D))$, all k, $\Rightarrow D$ is periodic. We assumed otherwise, so $\{w_k'\}_{k \in Z}$ is distinct, i.e. π is multiplicity free. Thus we are reduced to considering whether w_0' is an isolated point of $X^\#$ or not. But again, since $w_k'(A_n) \neq 0$ and $w_k'(B_n) \neq 0$, a simple argument shows that $w_{n_i}' \to w_0'$ if and only if $w_{n_i} \to w_0$. Putting w_{n_i} in canonical form, this is exactly the condition of the theorem.

PART III. UNILATERAL WEIGHTED SHIFTS

3.1. The generated C^*-algebra. If H is a separable Hilbert space, an operator W is a unilateral weighted shift means that for some orthonormal

basis $\{e_n\}_{n\in Z^+}$, $We_i = d_i e_{i+1}$, with $d_i \in C$. It is well known that such a W is irreducible if and only if each $d_i \neq 0$. We shall always assume this. Further, to unitary equivalence it may be assumed that $d_i > 0$ [15]. Then W has polar decomposition $W = S \cdot D$ where $D = \mathrm{diag}\{d_n\}_{n\in Z^+}$ and S is the unilateral shift.

If W has closed range, i.e. $\{d_i\}_{i\in Z^+}$ is bounded away from zero or alternatively D is invertible, we can apply Theorem 1.4.1. With the notation introduced there, I_S is generated by $S^*S - SS^*$ a projection of rank one. Any ideal in an irreducible C^*-algebra is irreducible, so $I_S = C(H)$ [2]. Now

$$\mathfrak{D} = C^*(\{\phi^n(D), S^{*(n)}S^{(n)}\}_{n\in Z}) = C^*(\{\phi(D)\}_{n\in Z}),$$

since D has closed range, is commutative. Let X denote the spectrum of $\mathfrak{D}$. If q is the quotient map $L(H) \to L(H)/C(H)$, then $q(\mathfrak{D})$ is also commutative. Denote its spectrum by $q(X)$, and call it the essential diagonal spectrum of W. Now ϕ induces an automorphism of $q(\mathfrak{D})$, and hence a homeomorphism, also denoted by ϕ, of $q(X)$. Condition (C) is easily verified so it follows that

Theorem 3.1.1. *If W is an irreducible unilateral weighted shift with closed range, then $C^*(W)/C(H)$ is *-isomorphic to the covariance algebra $C^*(q(X), \phi)$.*

It is not difficult to do as in §2.3 and extend this result to certain types of shifts which do not necessarily have closed range. We omit the details.

Since $C^*(W)$ contains the irreducible ideal $C(H)$, it is known [2], [8] that every representation is a direct sum of two subrepresentations ρ_1 and ρ_2 with $\rho_1(C(H)) \neq 0$, and $\rho_2(C(H)) = 0$. Then ρ_1 must be equivalent to a multiple of the identity representation, so with the representations ρ_2 of $C^*(q(X), \phi)$ having been described in Part I, quite a good description of the representations of $C^*(W)$ is possible; in particular, a complete description in case W is G.C.R. Additionally, the conditions given in §1.3 characterize which $C^*(W)/C(H)$ and thence $C^*(W)$ are G.C.R. We must postpone the corresponding characterization for those shifts without closed range until later.

As was done for the bilateral operators, since $\mathfrak{D}$ consists of diagonal operators, define $w_n \in X$, $n = 0, 1, 2$, by $w_n(B) = b_n$ if $B = \mathrm{diag}\{b_n\}_{n\in Z^+}$. By the usual argument $\{w_n\}_{n\in Z^+}$ is dense in X.

Now $\phi\colon D \to SDS^*$ and $\phi^{-1}\colon D \to S^*DS$ are both continuous, linear, multiplicative maps of $\mathfrak{D}$ into itself, the former (1-1) and the latter onto. If the zero homomorphism is adjoined to X, then ϕ and ϕ^{-1} induce continuous

maps of X into X, whose action on $\{w_n\}_{n \in Z^+}$ is given by

$$\begin{aligned} \phi(w_n) &= w_{n-1} \quad \text{if } n > 0, \\ &= 0 \quad \text{if } n = 0 \end{aligned}$$

and

$$\phi^{-n}(w_0) = w_n \quad \text{if } n \geq 0.$$

Let $X_d = \{0, w_0, w_1, \dots\}$ and $X_l = X - X_d$. If $y \in X_l$, then $y = \lim w_{n_i} = \lim \phi^{-n_i}(w_0)$, so $\phi^{-1} \circ \phi(y) = y$ since $\phi \circ \phi^{-1} = \mathrm{id}_X$, and ϕ is seen to be a homeomorphism of X_l. Of course X_l is simply the essential diagonal spectrum $q(X)$. For, if $x = \lim w_{n_i}$, and $K \in \mathfrak{D}$ is compact, then $x(K) = \lim w_{n_i}(K) = 0$. So $X_l \subset q(X)$. But $q(X) \subset X$ is always true and clearly no point of X_d, except 0, is in $q(X)$. So we have $X_l = q(X)$.

We again consider the canonical form of the diagonal spectrum. Thus if $x \in X$, let $T(X)$ be as in (2.2). Then T is a homeomorphism of X into $\Pi_{-\infty}^{\infty} Y$ where Y is the spectrum of D. In particular $T(w_0) = (\dots, 0, 0, 0, \underline{d_0}, d_1, d_2, \dots)$ and $T(w_1) = (\dots, 0, 0, d_0, d_1, d_2, \dots)$, etc. and $T(X) = \overline{\{T(w_n)\}}_{n \in Z}$ $= T(X_d) + T(X_l)$, where + denotes disjoint union.

Theorem 3.1.2. *Let ϕ be a homeomorphism of a compact Hausdorff space Y. Then there exists a unilateral weighted shift W for which $C^*(W)/C(H)$ is naturally *-isomorphic to $C^*(Y, \phi)$ if and only if*

(i) *$\exists f \in C(Y, R)$ with $\{f \circ \phi^n\}_{n \in Z}$ separating points.*

(ii) *If T denotes the natural isomorphism*

$$T(x) \rightarrow (\dots, f(\phi(x)), f(x), f(\phi^{-n}(x)), \dots)$$

then $\exists \{d_i\}_{i \geq 0}$, bounded, $\subset R$, such that if $D = (\dots, 0, 0, 0, d_0, d_1, d_2, \dots)$ and χ is the backward shift, then $\overline{\{\chi^k(D)\}}_{k \geq 0} = T(Y) + \{\chi^k(D)\}_{k \geq 0}$.

Proof. The remarks preceding the theorem show the necessity. For sufficiency, if we choose $W = S \cdot (D + 2 \sup |d_i| \cdot 1)$, where S = unilateral shift, then the canonical form of X_l described above shows that $q(X) = X_l$ $= T(Y)$, and hence that $C^*(W)/C(H)$ is indeed *-isomorphic to $C^*(Y, \phi)$.

It is unfortunate that as even very simple examples show it is necessary in the above to consider points outside Y. There is a simpler condition that is sufficient.

Corollary. *For sufficiency* (ii) *may be replaced by*

(ii)′ *for some $x \in Y$, either $\{\phi^n(x)\}_{n \in Z^+}$ or $\{\phi^{-n}(x)\}_{n \in Z^+}$ is dense in Y.*

Proof. Just let $d_i = f(\phi^{-i}(x))$, $i \geq 0$, in case $\{\phi^n(x)\}_{n \in Z^+}$ is dense.

Among the many pairs (Y, ϕ) to which the corollary applies are those of Examples 2 and 3 of Part II.

3.2. **N.G.C.R. shifts.** As was previously remarked, an irreducible operator W is N.G.C.R. if and only if $C^*(W) \cap C(H) = \{0\}$.

Theorem 3.2.1. *If $W = S \cdot D$ is an irreducible unilateral weighted shift, then W is N.G.C.R. if and only if there exists $n_i \to +\infty$ such that $d_{n_i+k} \to$ if $k \geq 0$ and $d_{n_i+k} \to 0$ if $k < 0$.*

Proof. After the description of the canonical diagonal spectrum given in §3.1, the theorem is proven in exactly the same manner that Theorem 2.5.1 was established.

The condition is clearly not a vacuous one, so the existence of such shifts is established. The result may also be expressed as follows.

Theorem 3.2.2. *If W is an irreducible unilateral weighted shift, $W = S \cdot D$, and $\mathfrak{D} = C^*(\{\phi^n(D)\}_{n \in Z})$, then $C^*(W)$ is N.G.C.R. if and only if ϕ^{-1} is an isomorphism of $\mathfrak{D}$.*

Proof. kernel $\phi^{-1} = \{\mathrm{diag}\{b, 0, 0, 0, \dots\}\} \subset C(H)$, so ϕ^{-1} is 1-1 if and only if $\mathfrak{D} \cap C(H) = \{0\}$.

3.3. **Algebraic equivalence.** We want to consider when two irreducible unilateral weighted shifts are algebraically equivalent. Firstly

Theorem 3.3.1. *If W_1 is any irreducible unilateral weighted shift with closed range, then W_1 is algebraically equivalent to another shift W_2 if and only if $W_1^* W_1 = W_2^* W_2$.*

Proof. Notice that W_2 is not assumed to be irreducible. It is well known that W_1 is unitarily eqiuvalent to W_2 if and only if $W_1^*W_1 = W_2^*W_2$. Hence the sufficiency. For the necessity we must first show that W_2 is in fact necessarily irreducible. For this, suppose W_2 has polar decomposition $W_2 = S' \cdot D_2$. Since W_1 has closed range, the unilateral shift S belongs to $C^*(W)$, and it is a consequence of Lemma 1.4.1 that if π implements the algebraic equivalence, then $\pi(S) = S'$. Hence S' is an isometry. But W_2 is a unilateral weighted shift, so it must be that $S' = S$, i.e. W_2 is irreducible.

Now since $C^*(W_1) \supset C(H)$, and π is an irreducible representation, π must be unitarily implemented [2]. So the conclusion.

If W_1 is not of closed range, it is possible that W_2 be reducible.

Example 4. Let $W_1 = S \cdot D_1$ where

$$D_1 = \mathrm{diag}\{1, 1, 1/2, 1, 1, 1/3, 1, 1, 1/4, \ldots\}$$

then W_1 has a three dimensional representation as the operator with matrix

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Thus W_1 has a faithful representation as the reducible shift with weights $\{1, 1, 0, 1, 1, \frac{1}{2}, \ldots\}$.

It is true for any irreducible operator that is not N.G.C.R., that every faithful irreducible representation is unitarily implemented. But we have seen that shifts with nonclosed range may be N.G.C.R. However, as has been true throughout most of this part, the unilateral case is not markedly different from the bilateral.

Theorem 3.3.2. *If W_1 and W_2 are irreducible unilateral weighted shifts, then W_1 and W_2 are algebraically equivalent if and only if $X_{1c} = X_{2c}$ (where X_{ic} denotes the canonical diagonal spectrum defined previously in* §3.1).

Proof. Let ρ be the representation implementing the algebraic equivalence. Extend ρ to a representation ρ' of $C^*(W, S)$ on some $H_{\pi'} \supset H_\pi$ in the usual way [8]. If ρ is not unitarily implemented, then certainly both W_1 and W_2 must be N.G.C.R. Let $\widetilde{W}_1$ denote W_1 with all the weights increased by one. Then ρ' is an irreducible representation of $C^*(\widetilde{W}_1) \approx C^*(\widetilde{X}_{1c}, \chi)$, for which $\rho'(W_1)$ is reducible and $\rho'(W_1)|_{H_\pi} = W_2$. This cannot occur unless ρ' is a transitive representation based on the orbit of a point in $\widetilde{X}_{1c}$ of the form $\{\ldots, a_{-2}, a_{-1}, 1, 1 + a_1, 1 + a_2, \ldots\}$ where $\{a_n\}_{n \in Z^-}$ are irrelevant and W_2 has weights $\{a_n\}_{n \in Z^+}$. But W_2 is N.G.C.R. so by Theorem 3.2.1 some sequence of translates converges to $\{\ldots, 1, 1, 1, 1 + a_1, 1 + a_2, 1 + a_3, \ldots\}$. This says that $X_{1c} \supset X_{2c}$, so by symmetry we are done.

3.4. **G.C.R. shifts.** For weighted shifts, either unilateral or bilateral, with nonzero weights and closed range, a characterization of those which are G.C.R. follows from §1.3. In [4], it is shown that any shift whose weights consist only of 0's and 1's is G.C.R. More generally, suppose $V = U \cdot D_1$ ($W = S \cdot D_2$) is a bilateral (unilateral) weighted shift. Let Y_1 (Y_2) be the subset of the diagonal spectrum X_1 (X_2) given by $Y_i = \{w \in X_i$: if $w(\phi^k(D)) = 0$ then $w(\phi^{k+n}(D)) = 0$, all $n \geq 0$ or all $n \leq 0\}$.

Theorem 3.4.1. *V (resp. W) is G.C.R. if and only if every point of Y_1 (Y_2) is discrete in its orbit.*

Proof. Since every irreducible representation of $C^*(W)$ is either unitarily equivalent to the identity representation or else is a representation of $C^*(W)/C(H)$ (these are not mutually exclusive) the argument for the unilateral case reduces to the following for the bilateral case.

To every point of X_1 there corresponds the transitive irreducible representation of the covariance algebra $C^*(V, U)$ described in §1.1. If the point is in Y_1, then by restricting to $C^*(V)$ and possibly taking a subrepresentation, one obtains an irreducible representation ρ of V as a bilateral weighted shift or a unilateral weighted shift or the adjoint of the last. In any of these cases, if the point of U is not discrete in its orbit, then by Theorem 2.5.1 or Theorem 3.2.1 it follows that $\rho(C^*(V)) \cap C(H) = \{0\}$. Thus V is not G.C.R.

Conversely, suppose ρ is an irreducible representation of $C^*(V)$ for which $\rho(C^*(V)) \cap C(H_\rho) = \{0\}$. Since V is a centered operator, it follows that $\rho(V)$ is one [21]. If $\rho(V)$ and $\rho(V^*)$ have zero null space, then by Lemma 2.4.1, ρ extends to a representation ρ' of $C^*(V, U)$, which is a covariance algebra, also on H_ρ. If $\rho' = (\pi, L)$ has μ_π transitive, then $\rho(V)$ is an N.G.C.R. irreducible bilateral weighted shift, and so by Thoerem 2.5.1, there exists $y \in Y_1$, not discrete in its orbit. If μ_π is intransitive, then by Theorem 1.3.1 there exist points in X_1, in fact a set of nonzero measure of them, which are not discrete in their orbit. Then $N(\rho(V^*)) = \{0\}$ implies that some point in Y_1 has this property, i.e. $N((M_f L_F U_\phi)^*) = \{0\} \Rightarrow \mu_\pi(\{y\colon f(y) = 0\}) = 0$. Finally, if $\rho(V)$ is an irreducible unilateral shift or the adjoint of one, then an argument almost identical to that at the end of Theorem 3.3.2 gives the conclusion. By the decomposition of centered operators given in [21], the above are the only possibilities for $\rho(V)$.

REFERENCES

1. W.B.Arveson, *Analyticity in operator algebras*, Amer. J. Math. 89 (1967), 578–642. MR 36 #6946.

2. ———, *Representations of C^*-algebras*, Lecture Notes in Math., Springer-Verlag, New York (to appear).

3. W.B. Arveson and K.B. Josephson, *Operator algebras and measure preserving automorphisms*. II, J. Functional Analysis 4 (1969), 100–134. MR 40 #3322.

4. J.W. Bunce and J.A. Deddens, *C^*-algebras generated by weighted shifts*, Indiana Univ. Math. J. 23 (1973), 257–271.

5. L.A. Coburn, *The C^*-algebra generated by an isometry*, Bull. Amer. Math. Soc. 73 (1967), 722–726. MR 35 #4760.

6. L.A. Coburn and A. Lebow, *Algebraic theory of Fredholm operators*, J. Math. Mech. 15 (1966), 577–584. MR 33 #569.

7. H.O. Cordes, *On a class of C^*-algebras*, Math. Ann. 170 (1967), 283–313. MR 35 #749.

8. J. Dixmier, *Les C^*-algèbres et leurs représentations*, Cahiers scientifique, fasc. 29, Gauthier-Villars, Paris, 1964. MR 30 #1404.

9. S. Doplicher, D. Kastler and D.W. Robinson, *Covariance algebras in field theory and statistical mechanics*, Comm. Math. Phys. **3** (1966), 1–28, MR **34** #4930.

10. E.G. Effros and F. Hahn, *Locally compact transformation groups and C*-algebras*, Mem. Amer. Math. Soc. No. 75 (1967). MR 37 #2895.

11. E.G. Effros, *Transformation groups and C*-algebras*, Ann. of Math. (2) **81** (1965), 38–55. MR **30** #5175.

12. J.G. Glimm, *Locally compact transformation groups*, Trans. Amer. Math. Soc. **101** (1961), 124–138. MR **25** #146.

13. A. Guichardet, *Utilisation des sous-groupes distingués ouverts dans l'étude des représentations unitaires des groupes localement compacts*, Compositio Math. **17** (1965), 1–35. MR **32** #5787.

14. P.R. Halmos, *Lectures on ergodic theory*, Publ. Math. Soc. Japan, no. 3, Math. Soc. Japan, Tokyo, 1956; reprint, Chelsea, New York, 1960. MR **20** #3958; **22** #2677.

15. ———, *A Hilbert space problem book*, Van Nostrand, Princeton, N.J., 1967. MR **34** #8178.

16. P.R. Halmos and J. von Neumann, *Operator methods in classical mechanics*. II, Ann. of Math. (2) **43** (1942), 332–350. MR **4**, 14.

17. K. Hoffman and R. Kunze, *Linear algebra*, Prentice-Hall Math. Ser., Prentice-Hall, Englewood Cliffs, N.J., 1961. MR **23** #A3146.

18. R.L. Kelley, *Weighted shifts on Hilbert space*, Ph. D. Dissertation, University of Michigan, Ann Arbor, Mich., 1966.

19. L.H. Loomis, *An introduction to abstract harmonic analysis*, Van Nostrand, Princeton, N.J., 1953. MR **14**, 883.

20. G.W. Mackey, *Unitary representations of group extensions*. I, Acta Math. **99** (1958), 265–311. MR **20** #4789.

21. B.S. Morrel and P.S. Muhly, *Centered operators* (preprint).

22. S. K. Parrot, *Weighted translation operators*, Ph.D. Dissertation, University of Michigan, Ann Arbor, Mich., 1965.

23. M. Takesaki, *Covariant representations of C*-algebras and their locally compact transformation groups*, Acta Math. **119** (1967), 273–303.

24. G. Zeller-Meier, *Porduits croisés d'une C*-algèbre par un groupe d'automorphismes*, J. Math. Pures Appl. (9) **47** (1968), 101–239. MR **39** #3329.

DEPARTMENT OF MATHEMATICS, STATE UNIVERSITY OF NEW YORK, STONY BROOK, NEW YORK 11794

Current address: Department of Mathematics, Dalhousie University, Halifax, Nova Scotia, Canada.

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

A STABILITY THEOREM FOR MINIMUM EDGE GRAPHS WITH GIVEN ABSTRACT AUTOMORPHISM GROUP

BY

DONALD J. McCARTHY[1] AND LOUIS V. QUINTAS

ABSTRACT. Given a finite abstract group $\mathcal{G}$, whenever n is sufficiently large there exist graphs with n vertices and automorphism group isomorphic to $\mathcal{G}$. Let $e(\mathcal{G}, n)$ denote the minimum number of edges possible in such a graph. It is shown that for each $\mathcal{G}$ there always exists a graph M such that for n sufficiently large, $e(\mathcal{G}, n)$ is attained by adding to M a standard maximal component asymmetric forest. A characterization of the graph M is given, a formula for $e(\mathcal{G}, n)$ is obtained (for large n), and the minimum edge problem is re-examined in the light of these results.

1. Introduction. Throughout this paper, automorphism groups of graphs will be regarded as abstract groups rather than permutation groups. It is well known [1] that given any finite group $\mathcal{G}$, there always exists a graph whose automorphism group is isomorphic to $\mathcal{G}$. It is natural to consider the following problem: Given a finite group $\mathcal{G}$, for each positive integer n decide whether or not there exists a graph on n vertices having automorphism group isomorphic to $\mathcal{G}$, and if there do exist such graphs determine the minimum number $e(\mathcal{G}, n)$ of edges possible. For a survey of results leading to this and related problems see [4].

To date, this minimum edge problem has been completely solved only when $\mathcal{G}$ is the identity group, a symmetric group, a dihedral group, or the cyclic group of order 3 [13], [14], [5], [3]. See also [16]. An examination of these cases reveals the following pattern. For small values of n, the behavior of $e(\mathcal{G}, n)$ may be somewhat erratic, due to the fact that the graphs for which $e(\mathcal{G}, n)$ is attained may fluctuate wildly, and for sporadic values of n no such graphs may exist. But eventually the situation becomes more stable: for n sufficiently large $e(\mathcal{G}, n)$ is always defined and, indeed, is

---

Presented to the Society, August 30, 1972 under the title *A stability theorem for minimum edge graphs with given abstract group*; received by the editors April 17, 1973.

AMS (MOS) subject classifications (1970). Primary 05C25, 20B25; Secondary 05C35, 05C05, 20D99.

Key words and phrases. Automorphism groups of graphs, minimum edge problem.

(1) Work of the first author was supported by National Science Foundation Grant GP-31723.

attained by adding to some fixed graph M a certain standard asymmetric forest. The point of the present paper is to show that this stability phenomenon occurs in general, for an arbitrary group $\mathfrak{G}$. A precise statement of this result is given at the end of the next section.

The theorem obtained characterizes M as a semireduced $\mathfrak{G}$-graph having minimum defect d_0 (see §2 for definitions) and a minimum number v_0 of vertices. It establishes, for large n, the formula $e(\mathfrak{G}, n) = n + d_0 - c$, where c denotes the number of components in a certain standard asymmetric forest on $n - v_0$ vertices. The basic properties of these forests are given in §3, and the proof of the stability theorem is given in §4.

Some consequences of the theorem are examined in §5; the minimum edge problem is reformulated, and observations are made regarding the stability graph M. These observations are applied, in §6, to the case where $\mathfrak{G}$ is commutative, and the cyclic case is treated in detail.

2. **Preliminaries and notation.** The bulk of the graph-theoretic terminology employed is in rough conformity with common usage [7], [11]. Throughout, all graphs are finite and undirected, without loops or multiple edges. Note that we permit the empty graph $\emptyset$, which has no vertices or edges.

For any graph G let $v(G)$, $e(G)$, $c(G)$ denote, respectively, the number of vertices, edges and connected components of G. By convention, $\emptyset$ is not a component of G. The *cycle rank* of G will be denoted by $r(G)$ and is given by $r = e - v + c$. We introduce the *defect* $d(G)$ defined by $d = r - c$.

An automorphism of G is a permutation of the set of vertices of G which preserves adjacency. The collection of all automorphisms of G foms a group which will be denoted by $\mathfrak{Aut}(G)$. The notation $\mathfrak{Aut}(G) \cong \mathfrak{G}$ will mean that $\mathfrak{Aut}(G)$ is isomorphic (as an abstract group) to the group $\mathfrak{G}$; in this case we refer to G as a $\mathfrak{G}$-graph. The identity group with only one element will sometimes be denoted by $\mathfrak{id}$, and an $\mathfrak{id}$-graph is termed *asymmetric*. By convention, $\emptyset$ is asymmetric.

Throughout, all groups are finite. The symmetric group of order $m!$ is designated by $\mathfrak{S}_m$, and $\mathfrak{S}_m[\mathfrak{G}]$ denotes the wreath product of $\mathfrak{G}$ by $\mathfrak{S}_m$. This last group is defined concretely in [12] and in [9, §3], and an abstract description is given below. Begin by taking the direct product of m copies of $\mathfrak{G}$; say $\mathfrak{D} = \mathfrak{G}_1 \times \mathfrak{G}_2 \times \ldots \times \mathfrak{G}_m$ where $\mathfrak{G}_i \cong \mathfrak{G}$. Let $\mathfrak{S}_m$ act on $\mathfrak{D}$ by permuting the subscripts in the natural manner. Finally, $\mathfrak{S}_m[\mathfrak{G}]$ is the semidirect product of $\mathfrak{D}$ by $\mathfrak{S}_m$ using this action. For details see [10, Chapter 6] or [9, §2].

Wreath products arise naturally as automorphism groups of graphs all of whose components are isomophic [2]. In general, suppose $G = m_1G_1 + m_2G_2 + \ldots + m_kG_k$ where the G_i are connected, G_i is not isomorphic to G_j for $i \neq j$, the symbol + denotes disjoint union, and m_iG_i is the disjoint union of m_i copies of G_i. If $\mathcal{A}ut(G_i) \cong \mathcal{G}_i$ then $\mathcal{A}ut(G) \cong \mathcal{S}_{m_1}[\mathcal{G}_1] \times \mathcal{S}_{m_2}[\mathcal{G}_2] \times \ldots \times \mathcal{S}_{m_k}[\mathcal{G}_k]$.

If C is a connected nonempty graph, the *multiplicity* of C in a graph G is defined as the number of components of G which are isomorphic to C. It follows from the preceding remarks that in a nonempty asymmetric graph all components have multiplicity 1. A graph G is said to be *reduced* if G has no proper subgraph H which is a union of components of G such that $\mathcal{A}ut(H) \cong \mathcal{A}ut(G)$. Equivalently, G is reduced if and only if G has no asymmetric components of multiplicity 1. Observe that the empty graph is the only reduced asymmetric graph. It is clear that every graph G can be decomposed in a unique manner as $R + A$ where R is reduced, A is asymmetric and $\mathcal{A}ut(G) \cong \mathcal{A}ut(R)$.

Let $a(G)$ denote the number of nonisomorphic asymmetric components of G, and let $a_1(G)$ be the number of nonisomorphic asymmetric components of multiplicity > 1 in G. Observe that if C is an asymmetric component of multiplicity m in G, then $\mathcal{A}ut(G)$ has a direct factor isomorphic to $\mathcal{S}_m[id] \cong \mathcal{S}_m$. Using the uniqueness of the direct product decomposition of a finite group into indecomposable factors [6, p. 130] and the indecomposability of $\mathcal{S}_m$, it follows that $a_1(C)$ cannot exceed the number of nontrivial symmetric groups which appear as direct factors in the decomposition of $\mathcal{A}ut(G)$, and the multiplicities m_i of these asymmetric components must equal the degrees of the corresponding symmetric groups $\mathcal{S}_{m_i}$.

Let $s(\mathcal{G})$ denote the number of nontrivial direct factors in the standard decomposition of $\mathcal{G}$ (into directly indecomposable groups) which are isomorphic to symmetric groups. We have just remarked that $a_1(G) \leq s(\mathcal{G})$ for every $\mathcal{G}$-graph G. We say that a $\mathcal{G}$-graph is *semireduced* if $a(G) \leq s(\mathcal{G})$. It is easy to see that the semireduced $\mathcal{G}$-graphs are precisely the graphs $R + A$ where R is a reduced $\mathcal{G}$-graph and A is a graph having at most $s(\mathcal{G}) - a(R)$ components, each of which is asymmetric and of multiplicity 1 in A and not isomorphic to any component of R. In particular, every reduced $\mathcal{G}$-graph is semireduced. Note also that when $s(\mathcal{G}) = 0$, every semireduced $\mathcal{G}$-graph is reduced.

We are now prepared to state our main result.

Theorem. *Let $\mathcal{G}$ be any finite group. Let d_0 denote the minimum defect $d(S)$, where S ranges over all semireduced $\mathcal{G}$-graphs; among all such graphs*

which satisfy $d(S) = d_0$ *let* M *be one having the smallest number of vertices* v_0.

Then for n *sufficiently large,* $e(\mathcal{G}, n) = n + d_0 - c(Q_{m,s})$ *where* $m = n - v_0$ *and* $s = s(\mathcal{G})$. *Indeed,* M *can be chosen so that* $e(\mathcal{G}, n)$ *is attained by* $M + Q_{m,s}$.

In the above, the graph $Q_{m,s}$ is a certain standard asymmetric forest. It has m vertices and a maximum number of components none of which is isomorphic to any member of a fixed set of s trees. The graphs $Q_{m,s}$ are defined precisely and their properties are investigated in the following section.

3. **Standard asymmetric forests.** Throughout we shall assume that all nonisomorphic asymmetric trees have been enumerated in some fixed order, $T_1, T_2, T_3, \ldots$, subject only to the condition that the trees are listed in order of increasing number of vertices, i.e. $v(T_i) \leq v(T_{i+1})$.

Following Quintas [13], we define a graph Q_n having n vertices, obtained essentially by taking an initial segment of this standard list of asymmetric trees. More precisely, suppose c is maximal so that $v(T_1 + T_2 + \ldots + T_c) \leq n$. If equality holds here, take $Q_n = T_1 + T_2 + \ldots + T_c$; otherwise we modify this slightly, replacing T_c by the first tree T_k in the list having exactly $n - v(T_1 + T_2 + \ldots + T_{c-1})$ vertices. Clearly Q_n is asymmetric, and $c = c(Q_n)$ is the maximum number of components in any asymmetric forest on n vertices. Since asymmetric trees on n vertices exist except when $1 < n < 7$, we see that Q_n is defined for $n = 1$ and $n \geq 7$.

We now make a slight generalization of this construction. In what follows, $\mathfrak{T}$ will denote a finite collection of nonisomorphic asymmetric trees. The graph $Q_n(\mathfrak{T})$ is defined by precisely the same procedure as above, but after first deleting the elements of $\mathfrak{T}$ from our standard list of asymmetric trees. When $\mathfrak{T} = \emptyset$ we have $Q_n(\mathfrak{T}) = Q_n$, and letting $\mathfrak{T}_s = \{T_1, T_2, \ldots, T_s\}$ define $Q_{n,s} = Q_n(\mathfrak{T}_s)$. A graph G is said to be $\mathfrak{T}$-free if each element of $\mathfrak{T}$ has multiplicity zero in G.

Lemma 1. *If* A *is an asymmetric* $\mathfrak{T}$*-free graph then* $d(A) \geq d(Q_n(\mathfrak{T}))$ *whenever* $Q_n(\mathfrak{T})$ *is defined and* $n \geq v(A)$.

Proof. If C is a component of A then $d(C) = -1$ if C is a tree, and $d(C) \geq 0$ otherwise. Thus if c_0 denotes the number of components of A which are trees, we have $d(A) \geq -c_0$. Each of the c_0 trees which occur as components of A is isomorphic to exactly one term in our standard list of asymmetric trees and none lies in $\mathfrak{T}$. It follows that the total number of vertices in the sum of the first c_0 trees in the list obtained by deleting $\mathfrak{T}$ from the standard list is at most $v(A)$. Since $n \geq v(A)$, it is clear from the construction

that $Q_n(\mathfrak{T})$ has at least c_0 components. Thus $d(Q_n(\mathfrak{T})) = -c(Q_n(\mathfrak{T})) \leq -c_0 \leq d(A)$. □

Since $e = v + d$ in general, it follows from the above result that $Q_n(\mathfrak{T})$ is a minimum edge graph among all asymmetric $\mathfrak{T}$-free graphs on n vertices. When $\mathfrak{T} = \varnothing$ this yields $e(id, n) = e(Q_n)$ whenever Q_n is defined, as shown by Quintas [13].

Lemma 2. *Given* $\mathfrak{T}$ *and a fixed integer* $k \geq 0$, *we have*

$$-1 \leq d(Q_{n+k}(\mathfrak{T})) - d(Q_n(\mathfrak{T})) \leq 0$$

provided only that n *is sufficiently large. More precisely, there exists an integer* $N_k(\mathfrak{T})$ *so that whenever* $n \geq N_k(\mathfrak{T})$, $Q_n(\mathfrak{T})$ *exists and the above inequalities hold.*

Proof. Assume for the moment that $Q_n(\mathfrak{T})$ exists whenever n is sufficiently large. Let $x = d(Q_{n+k}(\mathfrak{T})) - d(Q_n(\mathfrak{T}))$. It is clear (e.g. from Lemma 1) that $x \leq 0$ whenever both $Q_{n+k}(\mathfrak{T})$ and $Q_n(\mathfrak{T})$ are defined. Observe that $-x$ is the number of components by which $Q_{n+k}(\mathfrak{T})$ exceeds $Q_n(\mathfrak{T})$. Thus if n is so large that $Q_n(\mathfrak{T})$ already includes all available trees on at most k vertices, then clearly $-x \leq 1$ as desired.

To complete the proof we need only obtain an integer $N_0(\mathfrak{T})$ such that $Q_n(\mathfrak{T})$ exists whenever $n \geq N_0(\mathfrak{T})$. Then $N_k(\mathfrak{T})$ can be taken as the larger of $N_0(\mathfrak{T})$ and $v(F_k(\mathfrak{T}))$, where $F_k(\mathfrak{T})$ is the sum of all nonisomorphic asymmetric $\mathfrak{T}$-free trees on at most k vertices. Suppose m is an integer such that whenever $n \geq m$ there exists a $\mathfrak{T}$-free tree on n vertices. Letting $F = F_{m-1}(\mathfrak{T})$ it is clear that we can take $N_0(\mathfrak{T}) = m + v(F)$. Finally, we observe that m can be any integer ≥ 7 which exceeds the maximum number of vertices of any tree in $\mathfrak{T}$. For this last, we need only note that there exists an asymmetric tree on n vertices whenever $n \geq 7$. □

In what follows, we let $N_{k,s} = N_k(\mathfrak{T}_s)$. It will also be important in the sequel to obtain an integer $N_0(t)$ such that $Q_n(\mathfrak{T})$ exists for $n \geq N_0(t)$ whenever $\mathfrak{T}$ contains at most t trees. To do this, assume $t > 0$ and suppose m_t is an integer such that the number of nonisomorphic asymmetric trees on m_t vertices exceeds t. Clearly $n \geq m_t$ guarantees the existence of a $\mathfrak{T}$-free tree on n vertices for all collections $\mathfrak{T}$ containing at most t trees. Thus if F denotes the sum of all nonisomorphic asymmetric trees having fewer than m_t vertices, we can take $N_0(t) = m_t + v(F)$.

The existence of m_t follows from the well-known fact that a_n, the number of nonisomorphic asymmetric trees on n vertices, grows arbitrarily large with n. (This can be seen as follows. For $n \geq 7$, start with a path of length $n-2$, with vertices labelled consecutively as $v_1, v_2, \ldots, v_{n-1}$. At vertex v_i

append a path of length 1 to obtain a tree G_i having n vertices. For $2 < i \leq (n-1)/2$ the G_i are asymmetric and mutually nonisomorphic. Thus $a_n \geq [(n-1)/2] - 2 = [(n-5)/2]$. Better information on a_n can be obtained from [8].)

4. **Proof of the stability theorem.** We begin by observing that for any group $\mathfrak{G}$, there exist $\mathfrak{G}$-graphs on n vertices whenever n is sufficiently large. For suppose G_0 is any $\mathfrak{G}$-graph; existence of G_0 is guaranteed by [1]. Let $\mathfrak{T}$ denote the set of nonisomorphic asymmetric trees having multiplicity > 0 in G_0, and consider $G_0 + Q_m(\mathfrak{T})$. This is a $\mathfrak{G}$-graph on $n = v(G_0) + m$ vertices, existing whenever $n \geq v(G_0) + N_0(\mathfrak{T})$.

We wish to show that for n sufficiently large, the minimum number of edges for all $\mathfrak{G}$-graphs on n vertices is given by

$$e_n = n + d_0 - c(Q_{m,s}) \tag{4.1}$$

and that e_n is attained by a graph of the form $M + Q_{m,s}$ as indicated in the statement of the theorem. We prove the latter part first.

Thus, suppose M is a semireduced $\mathfrak{G}$-graph attaining minimum defect d_0 and having minimum number of vertices v_0. Letting $s = s(\mathfrak{G})$ as defined in §2, consider the graph $M + Q_{m,s}$ on $n = v_0 + m$ vertices. This graph exists for $n \geq v_0 + N_{k,s}$ but need not be a $\mathfrak{G}$-graph, since an asymmetric component of M may be repeated in $Q_{m,s}$. In that case, however, in view of the minimality of v_0 any such component must have the same number of vertices as some element of $\mathfrak{T}_s$ not appearing in M. It is possible to replace all occurrences of such components in M by appropriate elements of $\mathfrak{T}_s$ to obtain a semireduced $\mathfrak{G}$-graph M_1 whose asymmetric components all come from $\mathfrak{T}_s$. We have $d(M_1) = d_0$, $v(M_1) = v_0$ and $G_n = M_1 + Q_{m,s}$ is a $\mathfrak{G}$-graph on n vertices with $e(G_n) = n + d(G_n) = e_n$ as desired.

To complete the proof of the theorem, suppose G is any $\mathfrak{G}$-graph on n vertices; we will show that $e(G) \geq e_n$ provided only that n is sufficiently large. Again let $s = s(\mathfrak{G})$, $m = n - v_0$ and write $G = R + A$ where R is a reduced $\mathfrak{G}$-graph and A is asymmetric. Decompose A as $A_s + A_*$ where A_s is the sum of all those components of A isomorphic to elements of $\mathfrak{T}_s$. Let $\mathfrak{T} = \mathfrak{T}_s \cup \mathfrak{T}(R)$, where $\mathfrak{T}(R)$ denotes the set of all asymmetric trees having multiplicity > 0 in R. Thus $G = (R + A_s) + A_*$ and A_* is $\mathfrak{T}$-free and asymmetric. We distinguish two cases, according as $R + A_s$ is semireduced or not.

Case 1. $R + A_s$ is semireduced. Observe that $e(G) - e_n = x + y$ where $x = d(R + A_s) - d_0$ and $y = d(A_*) - d(Q_{m,s})$. In the present case we have $x \geq 0$, and by Lemma 1, $y \geq 0$ provided $v_1 \geq v_0$, where $v_1 = v(R + A_s)$. Thus

assume $v_1 < v_0$. By the minimality of v_0 we have $x > 0$, hence need only show $y \geq -1$. Let $k = v_0 - v_1$ so that $v(A_*) = m + k$. Since $k > 0$, $Q_{m+k,s}$ is defined whenever n (and hence m) is sufficiently large, and since A_* is $\mathfrak{T}$-free we have $d(A_*) \geq d(Q_{m+k,s})$. Thus $y \geq d(Q_{m+k,s}) - d(Q_{m,s})$ and the desired result follows from Lemma 2. More precisely, we obtain $y \geq -1$ whenever $n \geq v_0 + N_{k,s}$ where $k = v_0 - v_1 \leq v_0$. So in any event, $x + y \geq 0$ whenever $n \geq v_0 + N_{v_0,s}$.

Case 2. $R + A_s$ is not semireduced. If we let $a(R + A_s) = s + t$ then $0 < t \leq s$ and it is clear that A_s has at least t components and that R has at least t nonisomorphic asymmetric components not lying in $\mathfrak{T}_s$. Let C_1, $C_2, \ldots, C_t$ be such a set of components of R, and let $D_1, D_2, \ldots, D_t$ be components of A_s. Let m_i be the multiplicity of C_i in R and, finally, let H be the graph obtained from $R + A_s$ by replacing $m_i - 1$ copies of C_i by copies of D_i. Thus if we let $C = \Sigma(m_i - 1)C_i$ and $D = \Sigma(m_i - 1)D_i$, we have $H = R + A_s - C + D$.

Observe that if Q is any asymmetric $\mathfrak{T}$-free graph and $G' = H + Q$, then G' is a $\mathfrak{G}$-graph and in the decomposition $G' = R' + A'_s + A'_*$, the C_i appear only in A'_*; hence $R' + A'_s$ is semireduced. By Case 1, we have $e(G') \geq e_n$, provided only that $n' = v(G')$ is sufficiently large. Thus if we can select Q so that $n = n'$ and $e(G) \geq e(G')$, the desired result follows. We shall see that such a selection is possible provided $v(C)$ is sufficiently large. In the remaining case, i.e. when $v(C)$ is bounded by an appropriate constant, we make a different selection of Q which does not yield $n' = n$. Nevertheless, the boundedness of $v(C)$ guarantees that n' is large whenever n is sufficiently large, and we are still able to obtain $e(G) \geq e_n$. The details are presented below.

Symbolically we may write $G - G' = C - D + A_* - Q$. If we let $k = v(C) - v(D) + v(A_*)$ then $n - n' = k - v(Q)$. If $v(C)$ is sufficiently large, $k \geq v(A_*)$ and $Q_k(\mathfrak{T})$ exists. Taking $Q = Q_k(\mathfrak{T})$ gives $n' = n$ and $e(G) \geq e(G')$ as desired. (Since here $e(G) - e(G') = d(G) - d(G') = r(C) + d(A_*) - d(Q)$ and, via Lemma 1, $d(A_*) \geq d(Q)$.) It is not difficult to obtain a crude bound K so that $v(C) \geq K$ guarantees the appropriate conditions on k. E.g. $v(C) - v(D) \geq N_0(2s)$ ensures the existence of $Q_k(\mathfrak{T})$. If we let

$$L = v(\mathfrak{T}_s) \cdot \sum_{1 \leq i \leq s} (m_i - 1)$$

then certainly $v(D) \leq L$; hence we can take $K = L + N_0(2s)$.

Now we turn to the case in which $v(C) < K$. Here we select $Q = A_*$, so that $G' = G - C + D$. Consider the bracketed terms in the decomposition of

$e(G) - e_n$ as $\{e(G) - e(G')\} + \{e(G') - e_{n'}\} + \{e_{n'} - e_n\}$. Since $c(G') = c(G)$, we have $e(G) - e(G') = n - n' + r(C)$. Also, $e_{n'} - e_n = n' - n + z$ where $z = d(Q_{m',s}) - d(Q_{m,s})$ and $m' = n' - v_0$. Thus $e(G) - e_n = \{e(G') - e_{n'}\} + r(C) + z$. Since $n - n' = v(C) - v(D)$, the assumption that $v(C) < K$ implies that $n' > n - K$. This guarantees that for n sufficiently large we have $e(G') \geq e_{n'}$. (E.g. this is assured if $n \geq K + v_0 + N_{v0,s}$ and in what follows we assume n is at least this large.)

So we need only show that $r(C) + z \geq 0$ provided n is big enough. If $v(C) \geq v(D)$ then $n \geq n'$, hence $z \geq 0$ via Lemma 1. Since always $r(C) \geq 0$, there is nothing further to prove here. Thus suppose $v(C) < v(D)$. We will show that this implies $r(C) > 0$, hence require only $z \geq -1$. By Lemma 2, this last condition is met whenever n is sufficiently large (e.g. whenever $n \geq v_0 + N_{l,s}$ for any integer $l \geq v(D) - v(C)$; in particular we may take $l = L$). To see that $r(C) > 0$, simply observe that $v(C_i) \geq v(D_i)$ whenever C_i is a tree (since D_i lies in $\mathfrak{T}_s$ but C_i does not). Thus the assumption that $v(C) < v(D)$ implies that C cannot be a forest.

Thus in all cases, $e(G) \geq e_n$ whenever n is sufficiently large. In fact, we have shown that this holds whenever $n \geq v_0 + L + N$ where N is the maximum of $N_0(2s)$, $N_{v0,s}$ and $N_{L,s}$. □

5. **Consequences of the main result.** It follows from the theorem just proved that for an arbitrary group $\mathcal{G}$, as n grows large $e(\mathcal{G}, n)$ grows at the same rate as $e(\mathrm{id}, n)$. Indeed, for all n sufficiently large we have

$$d_0 + s \leq e(\mathcal{G}, n) - e(\mathrm{id}, n) \leq d_0 + s + 1. \tag{5.1}$$

To see this, use the fact that for n large enough

$$e(\mathrm{id}, n) = e(Q_n) = n - c(Q_n) \tag{5.2}$$

and $e(\mathcal{G}, n)$ is given by (4.1) with $m = n - v_0$ and $s = s(\mathcal{G})$. Observe also that if $u_s = v(T_1 + T_2 + \ldots + T_s)$ and $n' = u_s + m$, then $Q_{n'} = T_1 + T_2 + \ldots + T_s + Q_{m,s}$, hence

$$e(\mathcal{G}, n) = n + d_0 + s - c(Q_{n'}). \tag{5.3}$$

Thus $e(\mathcal{G}, n) - e(\mathrm{id}, n) = d_0 + s + x$ where $x = c(Q_n) - c(Q_{n'})$. It is easy to see that $v_0 \geq u_s$, so that $n \geq n'$ and an application of Lemma 2 yields $0 \leq x \leq 1$ as desired.

In view of the above, the rate of growth of $c(Q_n)$ may be of some interest, as is the problem of reasonably explicit computation of these numbers. In this regard, we remark only that $c(Q_n)$ can be described as in [13] in terms

of a_k, the number of nonisomorphic asymmetric trees on k vertices. Let $A_j = 1a_1 + 2a_2 + \ldots + ja_j$ and let u be maximal subject to $A_u \leq n$. Then $c(Q_n) = a_1 + a_2 + \ldots + a_u + w$, where w is the greatest integer in $(n - A_u)/(u+1)$. Although no closed formula for a_k is known, the generating function for this sequence is obtained in [8].

Leaving aside the problem of actual computation of $c(Q_n)$, in view of our stability theorem the complete determination of $e(\mathfrak{G}, n)$ for all n can be separated into several distinct subproblems:

(i) determine the values of $s(\mathfrak{G})$, d_0 and v_0;

(ii) determine the point at which stability first occurs; i.e. find the smallest integer N_0 such that $e(\mathfrak{G}, n) = e_n$ for $n \geq N_0$;

(iii) investigate the behavior of $e(\mathfrak{G}, n)$ when $n < N_0$.

It is hoped that this reformulation may be helpful in future work on the minimum edge problem. The stability result obtained here has little direct bearing on (ii) and (iii), although it is possible that in particular instances some of the ideas used in the proof may be of assistance. With regard to (i), computation of $s(\mathfrak{G})$ is a purely algebraic problem, of course, but determination of d_0 and v_0 may involve both algebraic and graph theoretic techniques. Also, (i) is closely related to the problem of actually exhibiting a "stability graph" for $\mathfrak{G}$, that is, a semireduced $\mathfrak{G}$-graph M with minimum defect and which attains the minimum number of vertices among such graphs. We offer several observations on the structure of M.

While the stability graph for a given group $\mathfrak{G}$ is by no means uniquely determined, some uniformizing assumptions can be made. In order to attain the minimum defect d_0 among semireduced $\mathfrak{G}$-graphs, M must contain precisely $s = s(\mathfrak{G})$ nonisomorphic asymmetric components. Since we also require M to have the minimum number v_0 of vertices, we may take these components to be the trees $T_1, T_2, \ldots, T_s$ and if m_i denotes the multiplicity of T_i in M we can also assume that $m_i \geq m_{i+1}$ for $1 \leq i < s$.

If M is reduced (i.e. if $m_s > 1$), the m_i are just the degrees of the nontrivial symmetric direct factors of $\mathfrak{G}$, and we have

$$M = m_1T_1 + m_2T_2 + \cdots + m_sT_s + M_0$$

where M_0 is a reduced $\mathfrak{G}_0$-graph. Here $\mathfrak{G}_0$ is a direct factor of $\mathfrak{G}$ which is maximal subject to $s(\mathfrak{G}_0) = 0$. It is clear that M_0 is a stability graph for $\mathfrak{G}_0$, and it follows that M_0 has no asymmetric components.

If the stability graph M always turned out to be a reduced $\mathfrak{G}$-graph, we could confine attention to groups with no nontrivial symmetric direct factors, for in view of the above observation, the general situation would present no

additional difficulties. Unfortunately, this is not the case.

For example, if $\mathcal{S}_2$ appears as a direct factor of $\mathcal{G}$, then the stability graph M is definitely not reduced whenever $s(\mathcal{G}) > 1$. For suppose M were reduced. We assume that for $i \leq s$, T_i has multiplicity $m_i > 1$ in M, and $m_s = 2$. Consider the graph M_1 obtained from M by replacing one copy of T_s by a path P_2 of length 1. Then M_1 is a semireduced $\mathcal{G}$-graph having the same defect as M but with fewer vertices. This contradicts the minimality of v_0.

We remark that although M itself need not be reduced, there always exists a reduced $\mathcal{G}$-graph R so that $e(\mathcal{G}, n)$ is attained for large n by $R + Q_{n',t}$ where $n' = n - v(R)$ and $t = a(R)$. (To see this, just take R to be the reduced part of the stability graph M.) This fact would be much more interesting if we could give a direct characterization of R, or at least of $v(R)$ and possibly $a(R)$. In some instances, R may be described in exactly the same manner as M. More precisely, let d_1 be the minimum defect of all reduced $\mathcal{G}$-graphs, and among such graphs attaining d_1 let R_1 be one having the smallest number v_1 of vertices. For many groups $\mathcal{G}$ we have $R = R_1$. If this were always the case, the emphasis on semireduced graphs in the theorem would be misplaced. We give an example in which this is not so.

Let $\mathcal{G} = \mathcal{S}_3 \times \mathcal{S}_2$. We have $s(\mathcal{G}) = 2$ and it is easy to see that $d_1 = -5$, $v_1 = 17$ and $R_1 = 3T_1 + 2T_2$. However, $d_0 = -5$, $v_0 = 12$ and $M = 3T_1 + P_2 + T_2$ where P_2 is a path having two vertices. Thus $R = 3T_1 + P_2 \neq R_1$. More important, for infinitely many n, $e(\mathcal{G}, n)$ is not attained by $R_1 + Q_{n-17,2}$. For, letting $m = n - 12$ and $m' = n - 17$ we have $c(Q_{m,2}) > c(Q_{m',2})$ and hence $e(M + Q_{m,2}) < e(R_1 + Q_{m',2})$ whenever $n = v(M + F_t)$ where $T_1 + T_2 + F_t$ is the sum of the first t trees in the standard list.

We make one final observation. There always exists a graph M' such that for n sufficiently large, $e(\mathcal{G}, n)$ is attained by $M' + Q_{n'}$ where $n' = n - v_0 + u_s$ as in (5.3). To see this, simply select M so that T_i has multiplicity > 0 in M for $1 \leq i \leq s$. If $M = M' + T_1 + \ldots + T_s$ then $M + Q_{m,s} = M' + Q_{n'}$. Of course, M' will not generally be a $\mathcal{G}$-graph.

6. **Application to the commutative case.** A good general description can be given for the stability graph of an arbitrary commutative group. Assume in what follows that $\mathcal{G}$ is commutative. Let $s = s(\mathcal{G})$ and let $\mathcal{G}_0$ denote a maximal direct factor of $\mathcal{G}$ with $s(\mathcal{G}_0) = 0$. When $s = 0$, $\mathcal{G} = \mathcal{G}_0$. When $s > 0$, $\mathcal{G} = \mathcal{G}_s \times \mathcal{G}_0$ where $\mathcal{G}_s$ is the direct product of s copies of a cyclic group of order 2.

We will employ the following result whose proof is omitted.

Lemma 3. *Let C be a component of multiplicity m in some $\mathcal{G}$-graph, where $\mathcal{G}$ is commutative.*

(1) *If $m > 1$, then $m = 2$, C is asymmetric, and $\mathcal{G}$ has a direct factor of order 2 arising from C.*

(2) *If C is a tree, then $\mathfrak{Aut}(C)$ is either trivial or a direct product of cyclic groups of order 2.*

(3) *If C is unicyclic and $\mathfrak{Aut}(C)$ has no direct factors of order 2, then $\mathfrak{Aut}(C)$ is cyclic.*

Suppose now that H is a semireduced $\mathcal{G}$-graph. It follows immediately from (1) and (2) that $d(H) \geq -2s$. Moreover, if equality holds here then a good deal can be said regarding the structure of H. Assume $d(H) = -2s$ and $s > 0$. We can write $H = H_s + H_0$ where H_s has precisely $2s$ components each of which is a tree whose automorphism group has order ≤ 2, and each component of H_0 has defect 0, i.e. is unicyclic. Furthermore, to obtain $c(H_s) = 2s$ it is necessary that $a(H_s) = s$ and $\mathfrak{Aut}(H_s)$ has order 2^s. It follows that H_s is a semireduced $\mathcal{G}_s$-graph and H_0 a reduced $\mathcal{G}_0$-graph. Also, if C is a component of H_0 then C has multiplicity 1, is unicyclic and $\mathfrak{Aut}(C)$ is cyclic of order > 2. When $s = 0$, $H_0 = H$ and has the same structural properties as when $s > 0$.

Further insight into the structure of the components of H_0 is provided by the following observation. Let U be an asymmetric unicyclic graph, and $q > 2$. Let $U(q)$ denote the unicyclic graph on $q \cdot v(U)$ vertices obtained by taking the natural q-fold covering of U. (If we regard U as a necklace, then $U(q)$ is obtained by unclasping U, taking q copies of this unclasped necklace and joining their clasps in the most natural manner.) It can be shown that $\mathfrak{Aut}(U(q)) \cong \mathcal{C}_q$, the cyclic group of order q, and that every unicyclic $\mathcal{C}_q$-graph arises in this manner. A detailed proof will be provided elsewhere.

Finally, we show that $d_0 = -2s$ by exhibiting a semireduced $\mathcal{G}$-graph H with $d(H) = -2s$. Take $H = 2(T_1 + T_2 + \ldots + T_s) + H_0$ where H_0 is as follows. Decompose $\mathcal{G}_0$ as a direct product of cyclic groups in some manner, and let the orders of the factors in this decomposition be $q_1, q_2, \ldots, q_k$. Let $U_1, U_2, \ldots, U_k$ be asymmetric unicyclic graphs chosen so that whenever $q_i = q_j$ with $i \neq j$, U_i is not isomorphic to U_j. Take H_0 to be the sum of the graphs $U_i(q_i)$ for $1 \leq i \leq k$. Each component of H_0 has multiplicity 1, hence H_0 is a $\mathcal{G}_0$-graph. It is clear that H meets the desired conditions.

We can obtain a stability graph M for $\mathcal{G}$ by modifying the above construction

to attain a minimum number of vertices. In more detail, it is clear from the result on the structure of semireduced $\mathcal{G}$-graphs of minimum defect that it is sufficient to obtain stability graphs M_0, M_s for $\mathcal{G}_0$, $\mathcal{G}_s$ respectively. Setting $M = M_0$ when $s = 0$, and $M = M_s + M_0$ when $s > 0$ yields a stability graph M for $\mathcal{G}$.

When $s > 0$, M_s is obtained as follows. Consider all nonisomorphic trees having automorphism group of order ≤ 2, and list them in order of increasing number of vertices: $W_1, W_2, \ldots$. Take

$$M_s = T_1 + \cdots + T_s + W_1 + \cdots + W_s;$$

clearly M_s is a stability graph for $\mathcal{G}_s$.

It is easy to describe an explicit procedure for obtaining M_0, but we gloss over the details. The procedure can be summarized as follows. To each decomposition D of $\mathcal{G}_0$ as a direct product of cyclic groups we assign a standard minimum-vertex $\mathcal{G}_0$-graph M_D whose components are unicyclic and have the factors appearing in D as their automorphism groups. If D is such that $v(M_D)$ is a minimum, we take $M_0 = M_D$.

The explicit description of M_D has been omitted here, but is not hard to reconstruct. The numbers $v(M_D)$ can be described entirely in terms of the orders of the factors in the decomposition D and the sequence b_k. Here b_k denotes the number of nonisomorphic asymmetric unicyclic graphs on k vertices; the generating function for b_k is known [15]. There seem to be genuine subtleties involved, however, in deciding which decomposition D minimizes $v(M_D)$. For a specific group $\mathcal{G}_0$, this can be settled by straightforward computation, of course, but the outcome in the general case cannot readily be described in simple terms as yet. There is, however, an important special case in which the result is clear.

Theorem. *Let $q_1, q_2, \ldots, q_k$ denote the orders of the factors in the decomposition of $\mathcal{G}_0$ as a direct product of cyclic groups of prime-power order. If $q_i \neq q_j \neq 2$ for $i \neq j$ $(1 \leq i, j \leq k)$, then $M_0 = C_1 + C_2 + \ldots + C_k$ where $C_i = U(q_i)$ and U is the unicyclic asymmetric graph on 6 vertices. In particular, $v(M_0) = 6(q_1 + q_2 + \ldots + q_k)$.*

As an immediate corollary, we obtain the following explicit description of the stability graph M in the case when $\mathcal{G}$ is cyclic. Suppose $\mathcal{G}$ is cyclic of order $q = q_1 q_2 \cdots q_k$ where the q_i are powers of different primes. If $q \not\equiv 2 \bmod 4$, then $s(\mathcal{G}) = 0$, hence $M = M_0$ as in the theorem. In particular, $d_0 = 0$ and $v_0 = 6\overline{q}$ where $\overline{q} = q_1 + q_2 + \ldots + q_k$. If $q \equiv 2 \bmod 4$, then $s(\mathcal{G}) = 1$, hence $\mathcal{G}_0$ is cyclic of order $q/2$. We have $M = 2T_1 + M_0$ where

M_0 is the stability graph for $\mathcal{G}_0$, as above. In particular, $d_0 = -2$ and $v_0 = 2 + 6(\overline{q} - 2)$.

REFERENCES

1. R. Frucht, *Herstellung von Graphen mit vorgegebener abstrakter Gruppe*, Compositio Math. 6 (1938), 239–250.

2. ———, *On the groups of repeated graphs*, Bull. Amer. Math. Soc. 55 (1949), 418–420. MR 10, 615.

3. R. Frucht, A. Gewirtz and L. V. Quintas, *The least number of edges for graphs having automorphism group of order three*, Recent Trends in Graph Theory (Proc. Conf., New York, 1970), Lecture Notes in Math., vol. 186, Springer-Verlag, Berlin and New York, 1971, pp. 95–104. MR 43 #6119.

4. A. Gewirtz, A. Hill and L. V. Quintas, *Extremum problems concerning graphs and their groups*, Combinatorial Structures and Their Applications (Proc. Calgary Internat. Conf. Combinatorial Structures and Their Applications, Calgary, Alberta, 1969), Gordon and Breach, New York, 1970, pp. 103–109. MR 42 #108.

5. G. Haggard, *The least number of edges for graphs having dihedral automorphism group*, Discrete Math. 6 (1973), 53–78. MR 48 #157.

6. M. Hall, Jr., *The theory of groups*, Macmillan, New York, 1959. MR 21 #1996.

7. F. Harary, *Graph theory*, Addison-Wesley, Reading, Mass., 1969. MR 41 #1566.

8. F. Harary and G. Prins, *The number of homeomorphically irreducible trees, and other species*, Acta Math. 101 (1959), 141–162. MR 21 #653.

9. A. Kerber, *Representations of permutation groups*. I, Lecture Notes in Math., vol. 240, Springer-Verlag, Berlin and New York, 1971.

10. B. H. Neumann, *Lectures on topics in the theory of infinite groups*, Tata Institute of Fundamental Research, Bombay, 1960.

11. O. Ore, *The theory of graphs*, Amer. Math. Soc. Colloq. Publ., vol. 38, Amer. Math. Soc., Providence, R. I., 1962. MR 27 #740.

12. ———, *Theory of monomial groups*, Trans. Amer. Math. Soc. 51 (1942), 15–64. MR 3, 197.

13. L. V. Quintas, *Extrema concerning asymmetric graphs*, J. Combinatorial Theory 3 (1967), 57–82. MR 35 #2780.

14. ———, *The least number of edges for graphs having symmetric automorphism group*, J. Combinatorial Theory 5 (1968), 115–125. MR 37 #6211.

15. P. K. Stockmeyer, *The enumeration of groups with prescribed automorphism group*, Ph. D. Dissertation, University of Michigan, Ann Arbor, Mich., 1971.

16. G. Haggard, D. McCarthy and A. Wohlgemuth, *Extremal edge problems for graphs with given hyperoctahedral automorphism group* (to appear).

DEPARTMENT OF MATHEMATICS, ST. JOHN'S UNIVERSITY, JAMAICA, NEW YORK 11439

DEPARTMENT OF MATHEMATICS, PACE UNIVERSITY, NEW YORK, NEW YORK 10038

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

INDUCED AUTOMORPHISMS ON FRICKE CHARACTERS OF FREE GROUPS

BY

ROBERT D. HOROWITZ(1)

ABSTRACT. The term character in this paper will denote the character of a group element under a general or indeterminate representation of the group in the special linear group of 2×2 matrices with determinant 1; the properties of characters of this type were first studied by R. Fricke in the late nineteenth century. Theorem 1 determines the automorphisms of a free group which leave the characters invariant. In a previous paper it was shown that the character of each element in the free group F_n of finite rank n can be identified with an element of a certain quotient ring of the commutative ring of polynomials with integer coefficients in $2^n - 1$ indeterminates. It follows that any automorphism of F_n induces in a natural way an automorphism on this quotient ring. Corollary 1 shows that for $n \geq 3$ the group of induced automorphisms of F_n is isomorphic to the group of outer automorphism classes of F_n. The possibility is thus raised that the induced automorphisms may be useful in studying the structure of this group. Theorem 2 gives a characterization for the group of induced automorphisms of F_2 in terms of an invariant polynomial.

1. Introduction. The algebraic properties of group characters under representation in the two-dimensional special linear group were first studied by R. Fricke [1] in connection with problems in the theory of Riemann surfaces. Although Fricke was primarily concerned with analytic questions, his work has led to results of group theoretic interest. Further results on free groups and related results have recently been given by the author [2] and A. Whittemore [5], [6]. Fricke in [1] observed the existence of naturally induced automorphisms on the ring of polynomial expressions in the characters of the

Received by the editors October 8, 1973.

AMS (MOS) subject classifications (1970). Primary 20E05, 20F55; Secondary 14E05, 20H05, 20H10.

Key words and phrases. Automorphisms of a free group, character under a general or indeterminate representation in the group of 2×2 matrices with determinant 1, induced automorphisms on the ring of polynomial expressions with integer coefficients in the characters.

(1) This paper is an extension of results contained in the author's doctoral thesis written at New York University. A portion of this research was supported by a grant from the City University of New York. The author would like to express his appreciation to his thesis adviser Professor Wilhelm Magnus.

group elements arising from the automorphisms of the group. The present paper investigates the induced automorphisms on the characters of free groups and their relationship to the automorphisms of the groups.

2. **Preliminaries.** SL_2 will denote the special linear group of 2×2 matrices with determinant 1 over the real or complex numbers. If G is a group, χu will denote the character of the element $u \in G$ under a general or indeterminate representation of G in SL_2. By this we mean that any relation which we write among the characters of elements of G will be understood to hold identically for all possible representations in SL_2. If we let $\mathfrak{R}_G$ denote the set of all representations of G in SL_2, and let $\mathfrak{F}(\mathfrak{R}_G, C)$ denote the ring of functions from $\mathfrak{R}_G$ to the complex numbers with the usual addition and multiplication, then the symbol χu can be regarded as formally denoting the function in $\mathfrak{F}(\mathfrak{R}_G, C)$ which assigns to each representation $\rho \in \mathfrak{R}_G$ the character *trace* $\rho(u)$ of u under ρ. The relations [1, p. 338]

$$\chi u^{-1} = \chi u, \tag{1}$$

$$\chi uv = \chi u \chi v - \chi uv^{-1}, \tag{2}$$

hold for all $u, v \in G$, and can be readily verified from the corresponding relations among the traces of arbitrary matrices in SL_2. The statement of the following result is due to Fricke [1]. A proof is given in [2]:

The character of an arbitrary element u in the free group F_n on the n generators $a_1, a_2, \ldots, a_n$ can be written as a polynomial expression

$$\chi u = P(\chi a_1, \chi a_2, \ldots, \chi a_1 a_2, \ldots, \chi a_1 a_2 \cdots a_n) \tag{3}$$

with integer coefficients in the $2^n - 1$ characters

$$\chi a_{i_1} a_{i_2} \cdots a_{i_\nu} \tag{4}$$

where $1 \leq i_1 < i_2 < \cdots < i_\nu \leq n$, $1 \leq \nu \leq n$.

The polynomial expression (3) is obtained by repeated application of the formulas (1), (2) to the freely reduced word representing u.

The following two lemmas will be used in the next section.

Lemma 1. *Let F be a free group on two or more generators $a, b, \cdots$. If $u \in F$ is such that $au^{-1}bu$ is conjugate to ab (or ba), then $u = b^l a^m$ for some integers l, m.*

The proof of Lemma 1 is a standard cancellation argument. Let U be the freely reduced word representing u. Let $U = b^l V a^m$ where V is freely

reduced, V not beginning with a power of b nor ending with a power of a. Then $aV^{-1}bV = a^m(aU^{-1}bU)a^{-m}$ is conjugate to ab. But $aV^{-1}bV$ is cyclically reduced as written. Therefore V must be the empty word.

Lemma 2 [2, Theorem 7.1]. *Let u be an element of the free group F. If $\chi u = \chi g^m$ where g^m is a power of a primitive element g, then u is conjugate to g^m or g^{-m}.*

3. Automorphisms leaving the characters invariant. We shall say that an automorphism α of the free group F *leaves the characters of F invariant* if $\chi u = \chi\alpha(u)$ for all $u \in F$.

Theorem 1. *Let I denote the group of automorphisms of the free group F which leave the characters of F invariant. If F has infinite rank, or if F has finite rank greater than or equal to three, then I is the group of inner automorphisms of F. If F has rank two, then I is the group generated by the inner automorphisms of F together with the automorphism which maps the two generators of F onto their inverses. If F has rank one, then I consists of the two automorphisms of F.*

Proof. The only automorphisms of F_1 are the identity automorphism and the automorphism $u \to u^{-1}$ which clearly leaves the characters of F_1 invariant by (1). Therefore we may restrict our attention to free groups of rank greater than or equal to two. Let a, b be the two generators of F_2. By (3), (4) it follows that the character of any element in F_2 can be represented as a polynomial expression in the three characters $\chi a, \chi b, \chi ab$. The image of this element under the automorphism $a \to a^{-1}, b \to b^{-1}$ will be the same expression in $\chi a^{-1}, \chi b^{-1}, \chi a^{-1}b^{-1}$. Since $\chi a^{-1} = \chi a$, $\chi b^{-1} = \chi b$, $\chi a^{-1}b^{-1} = \chi ba = \chi ab$ by (1), it follows that the automorphism $a \to a^{-1}, b \to b^{-1}$ leaves the characters of F_2 invariant. Clearly any inner automorphism leaves the characters of F invariant. Suppose conversely that α is an automorphism of F which leaves the characters invariant. We shall show that α is an inner automorphism or, in the case of F_2, the composition of an inner automorphism with $a \to a^{-1}, b \to b^{-1}$. Let S be the set of free generators which define the words of F. Let $g \in S$. Since α leaves characters invariant, it follows by Lemma 2 that $\alpha(g)$ must be conjugate to g or g^{-1}. Therefore

$$\alpha(g) = u_g^{-1} g^{\epsilon(g)} u_g \tag{5}$$

for some $u_g \in F$ and $\epsilon(g) = \pm 1$. Let a be a fixed element in S. Set $u_a = v$ and $\epsilon(a) = \epsilon$. Then (5) becomes

$$\alpha(a) = v^{-1}a^{\epsilon}v \tag{6}$$

when $g = a$. Let $\sigma = \pm 1$, and let $g \in S$ with $g \neq a$. Then since ag^{σ} is a primitive element of F and α leaves characters invariant, Lemma 2 implies that

$$\alpha(ag^{\sigma}) = w^{-1}(ag^{\sigma})^{\eta} w \tag{7}$$

for some $w \in F$ and $\eta = \pm 1$. Since $\alpha(ag^{\sigma}) = \alpha(a)[\alpha(g)]^{\sigma}$, we conclude from (7), (6) and (5) that

$$w^{-1}(ag^{\sigma})^{\eta} w = v^{-1} a^{\epsilon} v u_g^{-1} g^{\sigma\epsilon(g)} u_g. \tag{8}$$

The exponent sums on the left and right sides of (8) must be equal. Hence

$$(1 + \sigma)\eta = \epsilon + \sigma\epsilon(g). \tag{9}$$

Alternately setting $\sigma = -1$ and $\sigma = +1$ in (9) we obtain

$$\epsilon(g) = \epsilon \tag{10}$$

and $\eta = \epsilon$. If we set $\sigma = +1$ in (8), substitute $\eta = \epsilon(g) = \epsilon$, multiply on the left by v and on the right by v^{-1}, we obtain

$$vw^{-1}(ag)^{\epsilon} wv^{-1} = a^{\epsilon} v u_g^{-1} g^{\epsilon} u_g v^{-1}. \tag{11}$$

Since $\epsilon = \pm 1$, a^{ϵ} and g^{ϵ} are a pair of primitive elements for F. Therefore Lemma 1 implies that $u_g v^{-1} = g^{\epsilon l(g)} a^{\epsilon m(g)}$ for some integers $l(g)$, $m(g)$ depending on $g \in S$. If we substitute (10) and $u_g = g^{\epsilon l(g)} a^{\epsilon m(g)} v$ in (5), we conclude that

$$\alpha(g) = v^{-1} a^{-\epsilon m(g)} g^{\epsilon} a^{\epsilon m(g)} v \tag{12}$$

for all $g \in S$ with $g \neq a$. Let $g, h \in S$ with $g \neq h$, $g, h \neq a$. Then gh is a primitive element of F, and by (12)

$$\alpha(gh) = \alpha(g)\alpha(h) = v^{-1} a^{-\epsilon m(g)} g^{\epsilon} a^{\epsilon m(g) - \epsilon m(h)} h^{\epsilon} a^{\epsilon m(h)} v.$$

If we take characters of both sides using the fact that α leaves characters invariant and conjugate elements have the same character, we obtain

$$\chi gh = \chi g^{\epsilon} a^{\epsilon m(g) - \epsilon m(h)} h^{\epsilon} a^{\epsilon m(h) - \epsilon m(g)}. \tag{13}$$

Now since gh is a primitive element, it follows by Lemma 2 that the argument of the right side must be conjugate to $(gh)^{\pm 1}$. But this is impossible unless $m(g) = m(h)$, for otherwise, since a, g, h are distinct generators of F, the right side of (13) is cyclically reduced and has four syllables. Therefore $m(g)$ has a common value m for all $g \in S$, $g \neq a$. Thus (12) becomes

$$\alpha(g) = v^{-1}a^{-\epsilon m}g^{\epsilon}a^{\epsilon m}v \tag{14}$$

for all $g \in S$ with $g \neq a$. (14) is also clearly valid for $g = a$, for in this case (14) is equivalent to (6). Thus (14) holds for all $g \in S$. If $\epsilon = +1$ in (14), we see that α acts as a conjugation by $a^m v$ on every generator $g \in S$, and consequently on every element of F. Therefore, if $\epsilon = +1$, α is an inner automorphism. Suppose that $\epsilon = -1$. Then (14) becomes

$$\alpha(g) = v^{-1}a^{m}g^{-1}a^{-m}v \tag{15}$$

for all $g \in S$. Let $g, h \in S$ with $g \neq h$, $g, h \neq a$. Then

$$\alpha(gah) = \alpha(g)\alpha(a)\alpha(h) = v^{-1}a^{m}g^{-1}a^{-1}h^{-1}a^{-m}v$$

by (15). If we take characters on both sides using the fact that α leaves characters invariant and conjugate elements have the same character, we obtain $\chi gah = \chi g^{-1}a^{-1}h^{-1}$. However if g, h, a are distinct generators of F, this contradicts Lemma 2, as then gah is a primitive element of F, while $g^{-1}a^{-1}h^{-1}$ is clearly not conjugate to $(gah)^{\pm 1}$. Hence in this case there can be at most two generators in S. Since we are assuming F is not F_1 it follows that S has exactly two generators, and from (15) we see that α is the composition of an inner automorphism with the automorphism which maps each generator onto its inverse.

4. The induced automorphisms. Let $\mathcal{P}_n$ denote the commutative ring of polynomials with integer coefficients in the $2^n - 1$ indeterminates $x_{i_1 i_2 \ldots i_\nu}$. Let $\mathcal{I}_n$ denote the ideal consisting of all polynomials in $\mathcal{P}_n$ which vanish identically when the characters (4) are substituted for the corresponding indeterminates. Then the character χu of each element $u \in F_n$ can be identified with a unique element of the quotient ring $\mathcal{P}_n/\mathcal{I}_n$, the coset consisting of all polynomials $P \in \mathcal{P}_n$ which satisfy the right side of (3). Any automorphism α of F_n induces in a natural way a permutation $\chi u \to \chi\alpha(u)$ of the characters of F_n and, consequently, an automorphism on the ring generated by the characters together with the integer constants. Since the equivalence class $\{x_{i_1 i_2 \ldots i_\nu}\}$ of each indeterminate is identified with a character (4), it follows that the ring generated by the characters of F_n together with the integer constants is identified with the entire quotient ring $\mathcal{P}_n/\mathcal{I}_n$. The induced automorphism of F_n corresponding to α can consequently be regarded as an automorphism on the quotient ring $\mathcal{P}_n/\mathcal{I}_n$ given by

$$\{x_{i_1 i_2 \ldots i_\nu}\} \to \chi\alpha(a_{i_1}a_{i_2}\cdots a_{i_\nu}) \tag{16}$$

where the right side of (16) denotes the polynomial equivalence class in $\mathcal{P}_n/\mathcal{I}_n$

identified with the character. Let A_n denote the group of automorphisms of the free group F_n. Let I_n denote the subgroup of inner automorphisms, and let $J_n = A_n/I_n$ denote the quotient group of outer automorphism classes of F_n. It is readily verified that the induced automorphisms of F_n form a group $\mathfrak{A}_n$, and that the mapping which associates to each automorphism of F_n its corresponding induced automorphism on $\mathfrak{P}_n/\mathfrak{I}_n$ is a homomorphism from A_n to $\mathfrak{A}_n$. The kernel of this homomorphism is the group of all automorphisms in A_n which induce the identity automorphism. These are precisely the automorphisms of F_n which leave the characters invariant. This gives us the following result.

Corollary 1. *If $n \geq 3$, the groups $\mathfrak{A}_n$ and J_n are isomorphic.*

Corollary 1 raises the possibility that the induced automorphisms could be used to study the structure of the group J_n, $n \geq 3$. Results on the structure of the ideals $\mathfrak{I}_n$ for $n = 1, 2, 3, 4$ are given in [2] and [5].

5. The structure of $\mathfrak{A}_2$. The ideal $\mathfrak{I}_2$ is the zero ideal [2]. Thus the character χu of each $u \in F_2$ is given by a unique polynomial in $\mathfrak{P}_2$. Let a, b be the two generators of F_2. We set $x = x_1 = \chi a$, $y = x_2 = \chi b$, $z = x_{12} = \chi ab$. Then $\mathfrak{P}_2 = \mathbf{Z}[x, y, z]$ the commutative ring of polynomials with integer coefficients in x, y, z. The automorphisms

$$(17) \qquad \begin{array}{lll} a \to a^{-1} & a \to b & a \to ab \\ b \to b & b \to a & b \to b^{-1} \end{array}$$

together with the inner automorphisms generate A_2. (See e.g. [4, §4.5].) Consequently the induced automorphisms of (17) given by the right-hand column of (18) below constitute a generating set for $\mathfrak{A}_2$.

$$(18) \qquad \begin{array}{lll} \text{(i)} & a \to a^{-1} & x \to x \\ & b \to b & y \to y \\ & (ab \to a^{-1}b) & z \to xy - z \\ \text{(ii)} & a \to b & x \to y \\ & b \to a & y \to x \\ & (ab \to ba) & z \to z \\ \text{(iii)} & a \to ab & x \to z \\ & b \to b^{-1} & y \to y \\ & (ab \to a) & z \to x. \end{array}$$

(To see (18)(i) we observe that $\chi a^{-1}b = \chi ba^{-1} = \chi b \chi a - \chi ba = xy - z$ by

(2).) The following theorem essentially characterizes $\mathfrak{A}_2$ as a subgroup of the group of automorphisms of $\mathbf{Z}[x, y, z]$.

Theorem 2. *Let* $\mathfrak{A}_2^*$ *be the group of automorphisms of the ring* $\mathbf{Z}[x, y, z]$ *which keep invariant the polynomial*

$$C(x, y, z) = x^2 + y^2 + z^2 - xyz. \tag{19}$$

Then the automorphisms of $\mathfrak{A}_2$ *together with the two automorphisms*

$$\begin{array}{ll} x \longrightarrow -x & x \longrightarrow x \\ y \longrightarrow -y & y \longrightarrow -y \\ z \longrightarrow z & z \longrightarrow -z \end{array} \tag{20}$$

generate the automorphisms of $\mathfrak{A}_2^*$.

Remarks. The automorphisms (20) are not induced automorphisms because $-x$, $-y$, $-z$ cannot be characters. For under the representation which maps every element of F_2 onto the identity matrix of SL_2 we must have $\chi u = 2$ for all $u \in F_2$. Thus $\chi u \neq -\chi v$ for all $u, v \in F_2$. The polynomial (19) has the following significance. The relation

$$\chi aba^{-1}b^{-1} = x^2 + y^2 + z^2 - xyz - 2 \tag{21}$$

[1, p. 337, formula (8)] can be verified directly by matrix considerations or derived by applying formulas (1), (2). Since every automorphism in A_2 maps the commutator $aba^{-1}b^{-1}$ onto a conjugate of itself or its inverse, it follows that every induced automorphism must leave (21) and therefore (19) invariant.

Proof. Let $\mathfrak{A}_2^{**}$ be the group of automorphisms of $\mathbf{Z}[x, y, z]$ generated by $\mathfrak{A}_2$ together with the automorphisms (20). Clearly $\mathfrak{A}_2^{**} \subseteq \mathfrak{A}_2^*$. To complete the proof we must show $\mathfrak{A}_2^* = \mathfrak{A}_2^{**}$. Let

$$x \to P, \quad y \to Q, \quad z \to R \tag{22}$$

be an arbitrary automorphism in $\mathfrak{A}_2^*$ where P, Q, R are polynomials in $\mathbf{Z}[x, y, z]$. We wish to show that (22) lies in $\mathfrak{A}_2^{**}$. We can suppose without loss of generality that the degrees of P, Q, R are in ascending order

$$\deg P \leq \deg Q \leq \deg R. \tag{23}$$

For we see by (18)(ii), (iii) that the entire symmetric group on x, y, z is contained in $\mathfrak{A}_2$. Therefore we can apply a permutation to (22) to obtain an automorphism with degrees in ascending order. If this automorphism can be shown to be in $\mathfrak{A}_2^{**}$, it will then follow that the original automorphism (22) lay in $\mathfrak{A}_2^{**}$. Let

$$
\begin{aligned}
P &= P_p + P_{p-1} + \cdots + P_0, \qquad P_p \neq 0,\\
Q &= Q_q + Q_{q-1} + \cdots + Q_0, \qquad Q_q \neq 0, \qquad (24)\\
R &= R_r + R_{r-1} + \cdots + R_0, \qquad R_r \neq 0,
\end{aligned}
$$

where P_k, Q_k, R_k are homogeneous polynomials of degree k (i.e. P_k is the sum of all the terms of degree k in P; similarly for Q_k, R_k). If one of the polynomials P, Q, R consisted merely of a constant term, then (22) could not be composed with any mapping to produce the identity, and hence could not be an automorphism. Thus we must have $p, q, r \geq 1$ in (24). Since (22) keeps invariant the polynomial (19) we have

$$
-PQR + P^2 + Q^2 + R^2 = -xyz + x^2 + y^2 + z^2. \qquad (25)
$$

Suppose first that $p = q = r = 1$ in (24). If we then compare highest terms on the left and right of (25), we obtain $P_1Q_1R_1 = xyz$. Since x, y, z are irreducible, unique factorization implies that one of the polynomials P_1, Q_1, R_1 is $\pm x$, one is $\pm y$, and one is $\pm z$. We can suppose without loss of generality by composing (22) with a permutation as before that

$$
P_1 = c_1x, \quad Q_1 = c_2y, \quad R_1 = c_3z, \qquad (26)
$$

where $c_1, c_2, c_3 = \pm 1$ and $c_1c_2c_3 = 1$. If we substitute (26) in (24) and expand (25), we see that the term $c_2c_3P_1yz$ appears on the left while there is no term in yz on the right of (25). Therefore $P_0 = 0$. Similarly $Q_0, R_0 = 0$. Now the four possibilities for (22) are

$$
\begin{array}{llll}
x \to x & x \longrightarrow x & x \to -x & x \to -x\\
y \to y & y \to -y & y \longrightarrow y & y \to -y \qquad (27)\\
z \to z & z \to -z & z \to -z & z \longrightarrow z.
\end{array}
$$

All of these are elements of $\mathfrak{A}_2^{**}$ since they are all generated by the automorphisms (20). Now we proceed by induction on the maximum of the degrees of P, Q and R in (24). We may assume without loss of generality that $p \leq q \leq r$ so that this maximum is r. If we expand terms on the left side of (25) using (24), we obtain

$$
\begin{aligned}
-P_pQ_qR_r + \cdots + P_p^2 + \cdots + Q_q^2 + \cdots + R_r^2 + \cdots \qquad (28)\\
= -xyz + x^2 + y^2 + z^2,
\end{aligned}
$$

where the dots represent terms of lower order. Since the situation $p = q = r =$

1 has already been considered we may assume $r > 1$. Then the term $-P_pQ_qR_r$ in (28) is of degree at least 4. Since the right side of (28) has highest degree 3, it follows that all the terms of degree greater that 3 on the left side of (28) must cancel to zero. Thus there is no single term of highest degree on the left side of (28). We claim that $r = p + q$. For $r > p + q$ implies that $2r > 2p$, $2r > 2q$, $2r > p + q + r$, and then R_r^2 of degree $2r$ would be the highest term on the left side of (28) and the only term with this degree. Similarly $r < p + q$ implies that $p + q + r > 2r \geq 2p$, $2q$, and then $P_pQ_qR_r$ of degree $p + q + r$ would be the highest term on the left side of (28) and the only term with this degree, again leading to a contradiction. If $r = p + q$, then the terms of highest degree on the left side of (28) are $-P_pQ_qR_r + R_r^2 = 0$. Therefore

$$R_r = P_pQ_q. \tag{29}$$

Now consider the mapping

$$x \longrightarrow P, \quad y \longrightarrow Q, \quad z \longrightarrow PQ - R. \tag{30}$$

This mapping lies in $\mathfrak{A}_2^*$ since it is the composition of (22) with (18)(i). However (30) has highest degree less than r as deg $P = p < r$, deg $Q = q < r$ since $r = p + q$, and deg $PQ - R < r$ since the highest terms $P_pQ_q - R_r$ cancel because of (29). Therefore (30) lies in $\mathfrak{A}_2^{**}$ by the induction hypothesis. Now since (18)(i) is in $\mathfrak{A}_2^{**}$ it follows that the automorphism in (22) belongs to $\mathfrak{A}_2^{**}$ which completes the induction.

6. Remarks on $\mathfrak{A}_3$ and $\mathfrak{A}_4$. Unsolved problems. The question of the existence of analogous results to Theorem 2 for any of the groups $\mathfrak{A}_n$ where $n \geq 3$ remains open. We have shown in [2] that the ideal $\mathfrak{I}_3$ is a principal ideal generated by a polynomial of degree 4. One can readily obtain a generating set for the group $\mathfrak{A}_3$ by following the same procedure used for $\mathfrak{A}_2$. The induced automorphisms on $\mathfrak{P}_3/\mathfrak{I}_3$ thus obtained which generate $\mathfrak{A}_3$ are in turn seen to correspond to a set of automorphisms of $\mathfrak{P}_3$ which leave the polynomial generating $\mathfrak{I}_3$ invariant. In a communication to the author, Wilhelm Magnus has conjectured that an analogous result to Theorem 2 might be obtained for the group $\mathfrak{A}_3$ using the polynomial generator of $\mathfrak{I}_3$ as an invariant. The precise relationship between the automorphisms of $\mathfrak{A}_3$ and the automorphisms of $\mathfrak{P}_3$ which leave the ideal $\mathfrak{I}_3$ invariant awaits further investigation. Whittemore in [5] has given partial results on the structure of $\mathfrak{I}_4$, and has given a set of polynomials which collectively remain invariant under the automorphisms of $\mathfrak{A}_4$. These results appear to indicate that the structure of $\mathfrak{A}_n$ increases in complexity as n increases.

REFERENCES

1. R. Fricke and F. Klein, *Vorlesungen über die Theorie der automorphen Functionen*. Vol. 1, Teubner, Leipzig, 1897.

2. R. D. Horowitz, *Characters of free groups represented in the two dimensional special linear group*, Comm. Pure Appl. Math. 25 (1972), 635–649. MR 47 #3542.

3. ———, *Two dimensional special linear characters of free groups*, Ph. D. Thesis, New York University, New York, 1970.

4. W. Magnus, A. Karrass and D. Solitar, *Combinatorial group theory: Presentations of groups in terms of generators and relations*, Pure and Appl. Math., Vol. 13, Interscience, New York and London, 1966. MR 34 #7617.

5. A. Whittemore, *On special linear characters of free groups of rank $n \geq 4$*, Proc. Amer. Math. Soc. 40 (1973), 383–388. MR 48 #428.

6. ———, *On representations of the group of Listing's knot by subgroups of* SL(2, C), Proc. Amer. Math. Soc. 40 (1973), 378–382. MR 47 #5132.

DEPARTMENT OF MATHEMATICS, QUEENS COLLEGE (CUNY), FLUSHING, NEW YORK 11367

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

SOME ONE-SIDED THEOREMS ON THE TAIL DISTRIBUTION OF SAMPLE SUMS WITH APPLICATIONS TO THE LAST TIME AND LARGEST EXCESS OF BOUNDARY CROSSINGS

BY

Y. S. CHOW[1] AND T. L. LAI[2]

ABSTRACT. In this paper, we prove certain one-sided Paley-type inequalities and use them to study the convergence rates for the tail probabilities of sample sums. We then apply our results to find the limiting moments and the limiting distribution of the last time and the largest excess of boundary crossings for the sample sums, generalizing the results previously obtained by Robbins, Siegmund and Wendel. Certain one-sided limit theorems for delayed sums are also obtained and are applied to study the convergence rates of tail probabilities.

1. Introduction. Let $X_1, X_2, \ldots$ be i.i.d. random variables, and let $S_n = X_1 + \cdots + X_n$, $S_0 = X_0 = 0$. If $EX_1 = 0$, then for any $\epsilon > 0$, $P[|S_n| \geq \epsilon n]$ converges to 0 as $n \to \infty$. The rate at which the above convergence takes place, and more generally, the rate of convergence for $P[|S_n| \geq \epsilon n^\alpha]$, $\alpha > 1/2$, have been studied by a number of authors. In [1], Baum and Katz have proved that for any $p > 1/\alpha$, $\alpha > 1/2$, the following statements are equivalent:

$$\sum n^{p\alpha-2} P[|S_n| \geq \epsilon n^\alpha] < \infty \quad \text{for all } \epsilon > 0, \tag{1.1}$$

$$\sum n^{p\alpha-2} P\left[\sup_{k \geq n} |S_k/k^\alpha| \geq \epsilon\right] < \infty \quad \text{for all } \epsilon > 0, \tag{1.2}$$

$$E|X_1|^p < \infty, \quad \text{and for the case } \alpha \leq 1, \quad EX_1 = 0. \tag{1.3}$$

The analogous situation corresponding to the limiting case $\alpha = 1/2$ has been considered in [11], where it is proved that, for any $p > 2$, the following statements are equivalent:

Received by the editors December 12, 1973.

AMS (MOS) subject classifications (1970). Primary 60F10, 60F15, 60G50.

Key words and phrases. Convergence rates, Paley-type inequalities, last time, largest excess, limiting distribution, limiting moments, delayed sums.

[1] Research supported by the National Science Foundation under Grant NSF-GP-33570X at Columbia University.

[2] Research supported by the Office of Naval Research under Contract No. N00014-67-A-0108-0018 at Columbia University.

$$E|X_1|^p(\log^+|X_1|+1)^{-p/2}<\infty \quad \text{and} \quad EX_1=0, \tag{1.4}$$

$$\sum n^{p/2-2}P[|S_n| \geq \epsilon(n \log n)^{1/2}]<\infty \quad \text{for all large } \epsilon, \tag{1.5}$$

$$\sum n^{p/2-2}P\left[\sup_{k \geq n}|S_k/(k \log k)^{1/2}| \geq \epsilon\right]<\infty \quad \text{for all large } \epsilon. \tag{1.6}$$

It is natural to ask whether there are corresponding one-sided analogues of the above results. For example, if $1/2<\alpha<1$ and $X_1, X_2, \ldots$ are i.i.d. random variables with $EX_1=0$ and $E(X_1^+)^p<\infty$ for some $p>1/\alpha$, then is it always true that $\Sigma\, n^{p\alpha-2}P[S_n \geq \epsilon n^\alpha]<\infty$ for all $\epsilon>0$? The answer to this question turns out to be negative, as will be shown in §2 by a counterexample. However, if we also assume that $EX_1^2<\infty$, then the answer becomes affirmative. In fact, the following result has been established in [2]. Let $E|X_1|^r<\infty$ for some $1 \leq r \leq 2$, $E(X_1^+)^p<\infty$ for some $p \geq r$ and $EX_1=0$. If $\alpha r>1$, then

$$\sum n^{p\alpha-2}P\left[\max_{1 \leq k \leq n} S_k>\epsilon n^\alpha\right]<\infty \quad \text{for all } \epsilon>0. \tag{1.7}$$

The additional requirement $E|X_1|^r<\infty$ is a natural assumption, for without it, $P[S_n \geq \epsilon n^\alpha]$ may even converge to 1 under $EX_1=0$ and $E(X_1^+)^p<\infty$ for all p, as our counterexample shows. In §3, we shall obtain a sharper version of (1.7). A corresponding one-sided analogue of (1.5) and (1.6) under the assumption $EX_1=0$, $EX_1^2<\infty$ and $E(X_1^+)^p(\log(2+X_1^+))^{-p/2}<\infty$ will also be given in §4.

The series considered in (1.7) is closely related to the moments of the last time and of the largest excess of certain boundary crossings for the sequence X_n and for the sequence of partial sums S_n. More specifically, let us define

$$\begin{aligned} T(\epsilon,\alpha)&=\sup\{n \geq 1: S_n \geq \epsilon n^\alpha\} \quad (\sup \phi=0),\\ M(\epsilon,\alpha)&=\sup_{n \geq 0}(S_n-\epsilon n^\alpha),\\ T_1(\epsilon,\alpha)&=\sup\{n \geq 1: X_n \geq \epsilon n^\alpha\},\\ M_1(\epsilon,\alpha)&=\sup_{n \geq 0}(X_n-\epsilon n^\alpha). \end{aligned} \tag{1.8}$$

In §5, we shall consider the relations between the series in (1.7) and $E(T(\epsilon,\alpha))^{p\alpha-1}$, $E(T_1(\epsilon,\alpha))^{p\alpha-1}$, $E(M(\epsilon,\alpha))^{(p\alpha-1)/\alpha}$ and $E(M_1(\epsilon,\alpha))^{(p\alpha-1)/\alpha}$. Our results here extend those found in [3], [8], [9] and [15].

In §3, to sharpen the relation (1.7), we shall prove the following inequality: If $EX_1 = 0$ and $EX_1^2 < \infty$, then for $\alpha > 1/2$ and $p > 1/\alpha$,

$$(1.9)\quad \sum_{n=1}^{\infty} n^{p\alpha-2} P\left[\max_{1 \le k < n} S_k \ge n^{\alpha}\right] \le C_{p,\alpha}\{E(X_1^+)^p + (EX_1^2)^{(p\alpha-1)/(2\alpha-1)}\},$$

where $C_{p,\alpha}$ is a universal constant depending only on p and α. In fact, we shall derive a slightly more general inequality where we consider $E|X_1|^r$ in place of EX_1^2 for some $1 \le r \le 2$. The inequality (1.9) has some interesting applications in connection with the last time $T(\epsilon, \alpha)$ and the largest excess $M(\epsilon, \alpha)$ of boundary crossings. In §6, we shall show that as $\epsilon \downarrow 0$,

$$(1.10)\qquad \epsilon^{2/(2\alpha-1)} T(\epsilon, \alpha) \xrightarrow{\mathfrak{D}} T^*(\alpha),$$

$$(1.11)\qquad \epsilon^{1/(2\alpha-1)} M(\epsilon, \alpha) \xrightarrow{\mathfrak{D}} M^*(\alpha),$$

where "$\xrightarrow{\mathfrak{D}}$" denotes convergence in distribution, and

$$T^*(\alpha) = \sup\{t \ge 0 : W(t) \ge t^{\alpha}\}, \qquad M^*(\alpha) = \sup_{t \ge 0}(W(t) - t^{\alpha})$$

and $W(t)$, $t \ge 0$, is the standard Wiener process. Using the inequality (1.9), we easily obtain that if $E(X_1^+)^p < \infty$ for some $p > 2$, then

$$(1.12)\qquad \lim_{\epsilon \downarrow 0} \epsilon^{2(p\alpha-1)/(2\alpha-1)} E(T(\epsilon, \alpha))^{p\alpha-1} = E(T^*(\alpha))^{p\alpha-1},$$

$$(1.13)\qquad \lim_{\epsilon \downarrow 0} \epsilon^{(p\alpha-1)/\{\alpha(2\alpha-1)\}} E(M(\epsilon, \alpha))^{(p\alpha-1)/\alpha} = E(M^*(\alpha))^{(p\alpha-1)/\alpha}.$$

The inequality (1.9) enables us to simplify the proof and extend the result of Robbins, Siegmund and Wendel [15] and Kao [8] who in connection with certain statistical applications have considered the limiting relations (1.10), (1.11), (1.12) and (1.13) in the case $\alpha = 1$.

The one-sided inequality (1.9) obviously implies the corresponding two-sided result: If $EX_1 = 0$, $EX_1^2 < \infty$, $\alpha > 1/2$ and $p \ge 2$, then

$$(1.14)\quad \sum_{n=1}^{\infty} n^{p\alpha-2} P\left[\max_{1 \le k \le n} |S_k| \ge n^{\alpha}\right] \le C_{p,\alpha}\{E|X_1|^p + (EX_1^2)^{(p\alpha-1)/(2\alpha-1)}\}.$$

The above upper bound is sharp in the sense that a corresponding lower bound also holds:

$$(1.15)\quad 1 + \sum_{n=1}^{\infty} n^{p\alpha-2} P[|S_n| > n^{\alpha}] \ge B_{p,\alpha}\{E|X_1|^p + (EX_1^2)^{(p\alpha-1)/(2\alpha-1)}\}.$$

We shall refer to the inequalities (1.14) and (1.15) as Paley-type inequalities because of their resemblance to Paley's theorem which connects the type of integrability of a function with the rate of convergence of its Fourier coeffi-

cients (cf. [17, Vol. 2, p. 121]). The proof of (1.15) together with other related results and applications will be presented in another paper.

2. A counterexample. Let $0 < \delta < 2$ and define $\psi(x) = |x|^{-2}(\log|x|)^{-1-\delta}$ for $x \leq -c$, and let a, b, c be positive numbers such that $c \geq e$ and

$$a \int_{-\infty}^{-c} \psi(x)\,dx + b = 1, \qquad a \int_{-\infty}^{-c} x\psi(x)\,dx + b = 0.$$

Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with $P[X_1 = 1] = b$ and $P[X_1 \leq x] = a \int_{-\infty}^{x} \psi(x)\,dx$ for $x \leq -c$. Then $EX_1 = 0$ and $E(X_1^+)^p < \infty$ for all $p \geq 1$. Let $X_n' = X_n I_{[X_n \geq -n(\log n)^{-\delta/2}]}$, $S_n' = X_1' + \cdots + X_n'$. It is easy to see that

$$P[X_n \neq X_n' \text{ i.o.}] = 0, \tag{2.1}$$

$$ES_n' \sim (a/\delta) n (\log n)^{-\delta} \quad \text{as } n \to \infty, \tag{2.2}$$

$$\sigma(S_n') = (\operatorname{var} S_n')^{1/2} \sim (a/2)^{1/2} n (\log n)^{-1/2-3\delta/4} \quad \text{as } n \to \infty. \tag{2.3}$$

Since $\delta < 1/2 + 3\delta/4$, it follows from (2.2) and (2.3) that $\sigma(S_n') = o(ES_n')$. Therefore using the Tchebychev inequality, it is easy to see that $S_n'/(ES_n') \xrightarrow{P} 1$. This, together with (2.1), in turn implies that

$$(\delta/a) S_n (\log n)^{\delta}/n \xrightarrow{P} 1. \tag{2.4}$$

Hence $\lim_{n\to\infty} P[S_n > \epsilon n^{\alpha}] = 1$ for any $1 > \alpha > 1/2$, and so in contrast with (1.1), $\sum n^{-1} P[S_n \geq \epsilon n^{\alpha}] = \infty$. It is interesting to note that in the case $\alpha = 1$, $\sum n^{-1} P[S_n \geq \epsilon n] < \infty$ for all $\epsilon > 0$ by Spitzer's theorem [16].

We remark that our counterexample also gives a negative answer to the following question. The Marcinkiewicz-Zygmund strong law of large numbers states that if $X_1, X_2, \ldots$ are i.i.d. with $EX_1 = 0$ and $E|X_1|^p < \infty$ for some $1 \leq p < 2$, then $n^{-1/p} S_n \to 0$ a.e. It is natural to ask whether the one-sided analogue, i.e., $\limsup_{n\to\infty} n^{-1/p} S_n \leq 0$ a.e., would hold if $EX_1 = 0$ and $E(X_1^+)^p < \infty$. We note that this fails to hold in our example, since (2.4) implies that

$$\limsup_{n\to\infty} (\delta/a) S_n (\log n)^{\delta}/n \geq 1 \quad \text{a.e.} \tag{2.5}$$

3. A one-sided Paley-type inequality and its application to the convergence rate of tail probabilities. Let $X_1, X_2, \ldots$ be a sequence of random variables. Henceforth we shall use the following notation: For any real numbers t, $r \geq 1$,

$$(3.1)\qquad \begin{gathered} S_t = \sum_{i=1}^{[t]} X_i, \quad \overline{S}_t = \max_{1\le j\le t} S_j, \quad \overline{X}_t = \max_{1\le j\le t} X_j, \\ S_{r,t} = \sum_{r<j\le[r]+t} X_j, \quad \overline{S}_{r,t} = \max_{1\le j\le t} S_{r,j}, \quad S_0 = \overline{S}_0 = X_0 = \overline{X}_0 = 0. \end{gathered}$$

We first prove an inequality of which (1.9) is a special case.

Theorem 1. *Suppose $X_1, X_2, \ldots$ are i.i.d. with $EX_1 = 0$ and $E|X_1|^r<\infty$ for some $1 \le r \le 2$. Let $\alpha > 1/r$ and $p > 1/\alpha$. Then there exists a universal constant $C_{p,\alpha,r} > 0$ depending only on p, α and r such that*

$$\sum_{n=1}^{\infty} n^{p\alpha-2} P[\overline{S}_n \ge n^\alpha] \le C_{p,\alpha,r}\{E(X_1^+)^p + (E|X_1|^r)^{(p\alpha-1)/(r\alpha-1)}\}.$$

Proof. Let k be the smallest positive integer $> (p\alpha - 1)/(r\alpha - 1)$. We note that

$$(3.2)\qquad \begin{aligned} P[\overline{S}_n \ge n^\alpha] &\le P[\overline{X}_n > n^\alpha/(2k)] + P[\overline{S}_n \ge n^\alpha, \overline{X}_n \le n^\alpha/(2k)] \\ &\le nP[X_1 > n^\alpha/(2k)] + P[\overline{S}_n \ge n^\alpha, \overline{X}_n \le n^\alpha/(2k)]. \end{aligned}$$

For each fixed n, define

$$\tau_1 = \tau_1^{(n)} = \inf\{j \ge 1:\ S_j \ge n^\alpha/(2k)\},$$

$$\tau_2 = \tau_2^{(n)} = \inf\{j \ge 1:\ S_{\tau_1+j} - S_{\tau_1} \ge n^\alpha/(2k)\}, \text{ etc.}$$

Without loss of generality, we can assume that $E|X_1| \neq 0$. Since $EX_1 = 0$, it follows from the Chung-Fuchs theorem that $P[\tau_1 < \infty] = 1$. Also $\tau_1, \tau_2, \ldots$ are i.i.d. random variables. Hence

$$(3.3)\qquad \begin{aligned} P[\overline{S}_n \ge n^\alpha, \overline{X}_n \le n^\alpha/(2k)] &\le P[\tau_1 + \cdots + \tau_k \le n] \\ &\le P^k[\tau_1 \le n] = P^k[\overline{S}_n \ge n^\alpha/(2k)]. \end{aligned}$$

Now there exist positive constant A_r and B_r depending only on r such that

$$(3.4)\qquad \begin{aligned} P[\overline{S}_n \ge n^\alpha/(2k)] &\le A_r(2k)^r n^{-r\alpha} E|S_n|^r \quad \text{(cf. [4, p. 317])} \\ &\le B_r(2k)^r n^{-(r\alpha-1)} E|X_1|^r. \end{aligned}$$

The last inequality above follows from the Marcinkiewicz-Zygmund inequalities (cf. [13]). Letting $\lambda = k - (p\alpha - 2)/(r\alpha - 1)$, we have $\lambda > 1/(r\alpha - 1)$ and it follows from (3.4) that if $E|X_1|^r \ge 1$, then

$$\sum_{n^{r\alpha-1}>E|X_1|^r} n^{p\alpha-2}P^k[\bar{S}_n \geq n^{\alpha}/(2k)] \leq \xi_{p,\alpha,r}(E|X_1|^r)^k \sum_{n^{r\alpha-1}>E|X_1|^r} n^{-\lambda(r\alpha-1)}$$

$$\leq \zeta_{p,\alpha,r}(E|X_1|^r)^{k-\lambda+(1/(r\alpha-1))} \tag{3.5}$$

$$= \zeta_{p,\alpha,r}(E|X_1|^r)^{(p\alpha-1)/(r\alpha-1)}.$$

Also obviously

$$\sum_{n^{r\alpha-1}\leq E|X_1|^r} n^{p\alpha-2}P^k[\bar{S}_n \geq n^{\alpha}/(2k)] \leq \sum_{n^{r\alpha-1}\leq E|X_1|^r} n^{p\alpha-2} \tag{3.6}$$

$$\leq \eta_{p,\alpha}(E|X_1|^r)^{(p\alpha-1)/(r\alpha-1)};$$

$$\sum_{n=1}^{\infty} n^{p\alpha-1}P[X_1 \geq n^{\alpha}/(2k)] \leq \theta_{p,\alpha,r}E(X_1^+)^p. \tag{3.7}$$

Using (3.2), (3.3), (3.5), (3.6) and (3.7), we then obtain the desired conclusion for the case $E|X_1|^r \geq 1$.

If $E|X_1|^r < 1$, then it follows from (3.4) that

$$\sum_{n=1}^{\infty} n^{p\alpha-2}P^k[\bar{S}_n \geq n^{\alpha}/(2k)] \leq \gamma_{p,\alpha,r}(E|X_1|^r)^k \sum_{n=1}^{\infty} n^{-\lambda(r\alpha-1)} \tag{3.8}$$

$$\leq \rho_{p,\alpha,r}(E|X_1|^r)^{(p\alpha-1)/(r\alpha-1)}.$$

Hence the desired conclusion also follows in this case.

Corollary. *Suppose $X_1, X_2, \ldots$ are i.i.d., $EX_1 = 0$ and $E|X_1|^r < \infty$ for some $1 \leq r \leq 2$. Let $\alpha > 1/r$ and $p > 1/\alpha$. Then* (1.7) *is equivalent to each of the following statements*:

$$E(X_1^+)^p < \infty, \tag{3.9}$$

$$\sum n^{p\alpha-2}P\left[\sup_{k\geq n} k^{-\alpha}S_k \geq \epsilon\right] < \infty \quad \textit{for all } \epsilon > 0. \tag{3.10}$$

$$\sum n^{p\alpha-2}P[S_n \geq \epsilon n^{\alpha}] < \infty \quad \textit{for some } \epsilon > 0. \tag{3.11}$$

Proof. Replacing X_i by X_i/ϵ in Theorem 1, it is easy to see from Theorem 1 that (3.9) $\Rightarrow$ (1.7). By Lemma 2 in §5 below, (1.7) $\Rightarrow$ (3.10). Obviously (3.10) $\Rightarrow$ (3.11). By an argument due to Erdös [5] (see also Lemma 3 below) we can prove that (3.11) $\Rightarrow$ (3.9).

The results in the above corollary have been partly established in [2] by different methods. In [2], the case $\alpha = 1/r$ and $p \geq r$ for $1 \leq r < 2$ have also been considered and it is proved that in this case, (3.9) still implies (1.7). The following one-sided theorem on the convergence rate of tail probabilities deals with the case $\alpha > 1$. In this case, when $\alpha p > 1$, it follows immediately from Theorem 3 of [1] and the fact that $S_n \leq X_1^+ + \cdots + X_n^+$, while the situation $\alpha p = 1$ can be proved by using Theorem 1 (iii) and Lemma 3 of [2].

Theorem 2. *Suppose* $X_1, X_2, \ldots$ *are i.i.d.*, $\alpha > 1$ *and* $\alpha p \geq 1$. *If* $E(X_1^+)^p < \infty$, *then* (1.7) *holds, and consequently* (3.10) *also holds when* $\alpha p > 1$.

We remark that in Theorem 2, the relation (1.7) does not necessarily imply $E(X_1^+)^p < \infty$. To give an example, let $0 < p < 1$, $\alpha > 1/p$, $\gamma > p\alpha^2/(p\alpha - 1)$. Setting $q = p\alpha/\{\gamma(p\alpha - 1)\}$, we have $0 < q < p$. Choose $0 < \nu < 1$ such that $\gamma < \{\nu(p\alpha - 1)\}^{-1}$. Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $E(X_1^+)^q < \infty$, $E(X_1^+)^p = \infty$ and X_1^- has the stable distribution with exponent ν, i.e., the Laplace transform of X_1^- is given by $E \exp(-\lambda X_1^-) = \exp(-\lambda^\nu)$, $\lambda > 0$. Since $E(X_1^+)^q < \infty$, it follows from the Marcinkiewicz-Zygmund strong law of large numbers that

$$(3.12) \qquad \lim_{n\to\infty} n^{-1/q} \max_{j\leq n} (X_1^+ + \cdots + X_j^+) = 0 \quad \text{a.e.}$$

In particular, (3.12) holds along the subsequence $[m^{p\alpha/(p\alpha-1)} + m^{1/(p\alpha-1)}]$, and so

$$(3.13) \qquad \lim_{m\to\infty} m^{-\gamma} \max_{j\leq m^{p\alpha/(p\alpha-1)}+m^{1/(p\alpha-1)}} (X_1^+ + \cdots + X_j^+) = 0 \quad \text{a.e.}$$

This in turn implies that

$$(3.14) \qquad \lim_{m\to\infty} m^{-\gamma} S^{(+)}_{m^{p\alpha/(p\alpha-1)},m^{1/(p\alpha-1)}} = 0 \quad \text{a.e.}$$

where we define $S^{(+)}_{r,t} = \Sigma_{r<j\leq[r]+t} X_j^+$, and $S^{(-)}_{r,t} = \Sigma_{r<j\leq[r]+t} X_j^-$. Now $j^{-1/\nu} S^{(-)}_{r,j}$ has the same distribution as X_1^-, and it is well known that $P[X_1^- \leq t] = o(\exp(-t^{-\nu}))$ as $t \downarrow 0$. Since $\gamma < \{\nu(p\alpha - 1)\}^{-1}$, it follows from the Borel-Cantelli lemma that

$$(3.15) \qquad \lim_{m\to\infty} m^{-\gamma} S^{(-)}_{m^{p\alpha/(p\alpha-1)},m^{1/(p\alpha-1)}} = \infty \quad \text{a.e.}$$

From (3.14) and (3.15), we obtain

$$(3.16) \qquad \lim_{m\to\infty} m^{-\gamma} S_{m^{p\alpha/(p\alpha-1)},m^{1/(p\alpha-1)}} = -\infty \quad \text{a.e.}$$

Since $\alpha/(p\alpha - 1) < \gamma$, (3.16) implies that

$$(3.17) \qquad \limsup_{m\to\infty} m^{-\alpha/(p\alpha-1)} \overline{S}_{m^{p\alpha/(p\alpha-1)},m^{1/(p\alpha-1)}} \leq 0 \quad \text{a.e.}$$

By Lemma 3 of [2], (3.17) is equivalent to (1.7). We remark that the above example is similar to the one given by Baum [18] in another context.

Thus we have seen that in Theorem 2, (1.7) does not necessarily imply $E(X_1^+)^p < \infty$. A sufficient condition which would guarantee this implication is $E|X_1|^{1/\alpha} < \infty$. Under this additional condition, $\lim_{n\to\infty} nP[X_1 \geq \epsilon n^\alpha] = 0$, and by the Marcinkiewicz-Zygmund strong law of large numbers, $\lim_{n\to\infty} P[S_n \geq \epsilon n^\alpha] = 0$. Hence it can then be shown by the Erdös method that, in this case, (1.7) implies $E(X_1^+)^p < \infty$.

4. One-sided limit theorems for delayed sums and their relation to the convergence rate of tail probabilities. The quantity $S_{r,t}$ defined in (3.1) is called a delayed sum for the sequence X_n (cf. [17, Vol. 1, p. 80]). In [2], the following strong law for delayed sums has been proved: If $X_1, X_2, \ldots$ are i.i.d. with $EX_1 = 0$ and $E|X_1|^p < \infty$ for some $p \geq 1$, then for every $0 < \beta < \min(1, 2/p)$,

$$\lim_{n\to\infty} n^{-1/p} \max_{1\leq j\leq n^\beta} |S_{n,j}| = 0 \quad \text{a.e.} \tag{4.1}$$

The corresponding one-sided limit theorem has also been obtained: If $E|X_1|^r < \infty$ for some $1 \leq r \leq 2$, $E(X_1^+)^p < \infty$ for some $p \geq r$ and $EX_1 = 0$, then for every $0 < \beta < r/p$ (or for every $0 < \beta \leq 1/p$ in the case $r = 1$),

$$\limsup_{n\to\infty} n^{-1/p} \overline{S}_{n,n^\beta} \leq 0 \quad \text{a.e.} \tag{4.2}$$

For $\alpha r > 1$, obviously (4.2) implies (3.17) which is in turn equivalent to (1.7). Thus based on the equivalence between (1.7) and (3.17) which holds for any $\alpha > 0$ and $\alpha p > 1$, we can prove theorems concerning the convergence rate of tail probabilities from the corresponding limit theorems for delayed sums.

We note that while β ranges from 0 to $\min(1, 2/p)$ in (4.1), the range of β in (4.2) is from 0 to r/p. Our example in §2 shows that we cannot extend the range of β in (4.2). In that example, $r = 1$ and we can take any $p > 1$ since X_1^+ is bounded. Now for any $0 < \beta < 1$, since S_{n,n^β} has the same distribution as S_{n^β}, it is easy to see from (2.4) that

$$\limsup_{n\to\infty} (\delta/a) S_{n,n^\beta} (\beta \log n)^\delta / n^\beta \geq 1 \quad \text{a.e.} \tag{4.3}$$

Therefore if $\beta > 1/p$, then $\limsup_{n\to\infty} n^{-1/p} S_{n,n^\beta} = \infty$ a.e. This shows that we cannot extend the range of β in (4.2) beyond r/p. It is interesting to note that in spite of (2.4), we have for any $0 < \beta < 1$,

$$\liminf_{n\to\infty} S_{n,n^\beta} (\log n)^\delta / n^\beta = -\infty \quad \text{a.e.} \tag{4.4}$$

To prove (4.4), let $Z_n = (X_{n+1} + \cdots + X_{n+[n^\beta]})(\log n)^\delta/n^\beta$, $Y_n = X_n(\log n)^\delta/n^\beta$. Suppose (4.4) is not true. Then by the zero-one law, there exists a constant c such that $\liminf_{n\to\infty}(Y_n + Z_n) \geq c$ a.e. Since Z_n is independent of $(Y_1, \ldots, Y_n)$ and $Z_n \xrightarrow{P} a/\delta$, it follows from Lemma 1 in [11] that $\liminf_{n\to\infty} Y_n \geq c - (a/\delta)$ a.e. But $\liminf_{n\to\infty} Y_n = -\infty$ since $EX_1(\log|X_1|)^\delta = -\infty$, and so we have a contradiction.

The proof of the equivalence of (1.4), (1.5) and (1.6) in [11] is based on the following analogue of the law of the iterated logarithm for delayed sums: If $X_1, X_2, \ldots$ are i.i.d. and $0 < \beta < 1$, then

$$EX_1 = 0, \quad EX_1^2 = \sigma^2 \quad \text{and} \quad E|X_1|^{2/\beta}(\log^+|X_1| + 1)^{-1/\beta} < \infty$$

(4.5)

$$\Leftrightarrow \limsup_{n\to\infty} \max_{1\leq j\leq n^\beta} |S_{n,j}|/\{2(1-\beta)n^\beta \log n\}^{1/2} = \sigma \quad \text{a.e.}$$

Theorem 3 below gives the one-sided analogue of (4.5).

Theorem 3. *Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $EX_1 = 0$, $EX_1^2 = \sigma^2(<\infty)$. Then for $0 < \beta < 1$, the following statements are equivalent:*

$$\int_{[X_1>e]} X_1^{2/\beta}(\log X_1)^{-1/\beta}\, dP < \infty, \tag{4.6}$$

$$\limsup_{n\to\infty} \overline{S}_{n,n^\beta}/\{2(1-\beta)n^\beta \log n\}^{1/2} \leq \sigma \quad a.e., \tag{4.7}$$

$$\limsup_{n\to\infty} S_{n,n^\beta}/\{2(1-\beta)n^\beta \log n\}^{1/2} < \infty \quad a.e. \tag{4.8}$$

Lemma 1. *Let X be a nonnegative random variable such that $Eg(X) < \infty$ for some nonnegative Borel function g. Then there exists a positive nondecreasing function ψ on $[0, \infty)$ such that $\lim_{t\to\infty}\psi(t) = \infty$, $\lim_{t\to\infty}\psi(t^p)/\psi(t) = 1$ for all $p > 0$ and $E\psi(X)g(X) < \infty$.*

Proof. Let F be the distribution function of X. Let $(n_k)_{k\geq 1}$ be a sequence of integers such that $n_1 \geq 2$, $n_{k+1} > 2^{n_k}$ and $\int_{[n_k,\infty)} g(t)\, dF(t) < 2^{-k}$. Let $n_0 = 0$ and define $\psi(t) = k^{1/2}$ for $n_{k-1} \leq t < n_k$. Obviously $\psi(t) \uparrow \infty$ as $t \uparrow \infty$ and

$$\int_0^\infty \psi(t)g(t)\, dF(t) = \sum_{k=1}^\infty \int_{[n_{k-1},n_k)} \psi(t)g(t)\, dF(t)$$

$$\leq \int_0^{n_1} g(t)\, dF(t) + \sum_{k=2}^\infty k^{1/2} 2^{-(k-1)} < \infty.$$

Proof of Theorem 3. First suppose that (4.6) holds. To prove (4.7), we shall use the idea of Hartman and Wintner [7] in truncating the random variables from below. By Lemma 1, we can choose a positive nondecreasing function ψ on $[0, \infty)$ such that $\lim_{t\to\infty} \psi(t) = \infty$, $\lim_{t\to\infty} \psi(t^p)/\psi(t) = 1$ for all $p > 0$ and $EX_1^2\psi(|X_1|) < \infty$. Without loss of generality, we can assume that $\sigma > 0$. Given $\delta > 0$, we pick an integer $k > 1$ such that $k - \beta k > 1$ and then choose $\epsilon > 0$ such that $\epsilon k < \delta$. Define

$$X_n^{(1)} = X_n I_{[-n^{\beta/2}(\log n)^{-1/2}(\psi(n))^{-1/2} \le X_n \le n^{\beta/2}/(\log n)]},$$

$$X_n^{(2)} = X_n I_{[X_n < -n^{\beta/2}(\log n)^{-1/2}(\psi(n))^{-1/2}]},$$

$$X_n^{(3)} = X_n I_{[n^{\beta/2}/(\log n) < X_n < \epsilon n^{\beta/2}(\log n)^{1/2}]},$$

$$X_n^{(4)} = X_n I_{[X_n \ge \epsilon n^{\beta/2}(\log n)^{1/2}]}.$$

Let $U_n = X_n^{(1)} - EX_n^{(1)}$. Then $EU_n = 0$, $EU_n^2 = \sigma_n^2 \to \sigma^2$ and

$$|U_n| \le 2n^{\beta/2}\{(\log n)^{-1} + (\log n)^{-1/2}(\psi(n))^{-1/2}\} = \gamma_n = o(n^{\beta/2}(\log n)^{-1/2}).$$

Hence for $|t|\gamma_j \le 1$,

$$\exp\{t^2\sigma_j^2(1 - |t|\,\gamma_j)/2\} \le E\,\exp(tU_j) \le \exp\{t^2\sigma_j^2(1 + \tfrac{1}{2}|t|\gamma_j)/2\}$$

(cf. [12, p. 255]). Therefore by Theorem 1 of [11],

$$\limsup_{n\to\infty} \max_{1\le j\le n^\beta} |U_{n+1} + \cdots + U_{n+j}|/\{2(1-\beta)n^\beta \log n\}^{1/2} = \sigma \quad \text{a.e.} \tag{4.9}$$

We note that, since $EX_1 = 0$, we have for all large n,

$$E|X_n^{(1)}| = \left|\int_{[X_1 > n^{\beta/2}/(\log n)]} X_1\,dP + \int_{[X_1 < -n^{\beta/2}(\log n)^{-1/2}(\psi(n))^{-1/2}]} X_1\,dP\right|$$

$$\le 2\left\{(2/\beta)^{1/\beta} n^{\beta/2-1}(\log n)^{3/\beta-1}\int_{[X_1>e]} X_1^{2/\beta}(\log X_1)^{-1/\beta}\,dP\right.$$

$$\left. + n^{-\beta/2}(\log n)^{1/2}(\psi(n))^{-1/2}EX_1^2\psi(|X_1|)\right\}.$$

Therefore

$$\lim_{n\to\infty} \sum_{j=1}^{[n^\beta]} E|X_{n+j}^{(1)}|/\{n^\beta \log n\}^{1/2} = 0. \tag{4.10}$$

It is obvious that

$$(4.11)\qquad \limsup_{n\to\infty} \max_{1\le j\le n^\beta} (X_{n+1}^{(2)} + \cdots + X_{n+j}^{(2)})/\{n^\beta \log n\}^{1/2} \le 0 \quad \text{a.e.};$$

$$(4.12)\qquad \lim_{n\to\infty} \max_{1\le j\le n^\beta} (X_{n+1}^{(4)} + \cdots + X_{n+j}^{(4)})/\{n^\beta \log n\}^{1/2} = 0 \quad \text{a.e.}$$

Since $0 \le X_n^{(3)} < \epsilon n^{\beta/2}(\log n)^{1/2}$, it can be shown as in the proof of Theorem 2 in [11] that

$$P\left[\max_{1\le j\le n^\beta} (X_{n+1}^{(3)} + \cdots + X_{n+j}^{(3)}) \ge \delta (n^\beta \log n)^{1/2}\right]$$

$$\le P\left[\sum_{j=1}^{[n^\beta]} X_{n+j}^{(3)}/\{\epsilon n^{\beta/2}(\log n)^{1/2}\} \ge k\right]$$

$$= O(n^{\beta k}\{(\log n)^{3/\beta}/n\}^k).$$

Since $k - \beta k > 1$, an application of the Borel-Cantelli lemma gives

$$(4.13)\qquad \limsup_{n\to\infty} \max_{1\le j\le n^\beta} (X_{n+1}^{(3)} + \cdots + X_{n+j}^{(3)})/\{n^\beta \log n\}^{1/2} \le \delta \quad \text{a.e.}$$

As δ is arbitrary, we obtain (4.7) from (4.9), (4.10), (4.11), (4.12) and (4.13).

Obviously (4.7) implies (4.8). Now assume (4.8). By the zero-one law, there exists a finite constant c such that the lim sup in (4.8) is $\le c$ a.e. Define

$$Y_n = X_n/\{2(1-\beta)n^\beta \log n\}^{1/2},$$

$$Z_n = (X_{n+1} + \cdots + X_{n+[n^\beta]})/\{2(1-\beta)n^\beta \log n\}^{1/2}.$$

Since $EX_1 = 0$ and $EX_1^2 = \sigma^2$, $Z_n \xrightarrow{P} 0$. Obviously Z_n is independent of $(Y_1, \ldots, Y_n)$. Since $\limsup_{n\to\infty}(Y_n + Z_n) \le c$ a.e., we obtain using Lemma 1 of [11] that $\limsup_{n\to\infty} Y_n \le c$ a.e. From this, (4.6) follows easily.

By using a similar argument as the proof in [11] of the equivalence of (1.4), (1.5) and (1.6), we can obtain from Theorem 3 the following one-sided theorem on the tail distribution of sample sums.

Theorem 4. *Suppose $X_1, X_2, \ldots$ are i.i.d., $EX_1 = 0$, $EX_1^2 = \sigma^2(<\infty)$. Then for any $p > 2$, the following statements are equivalent:*

$$(4.14)\qquad \int_{[X_1>e]} X_1^p(\log X_1)^{-p/2}\,dP < \infty,$$

$$(4.15)\qquad \sum n^{p/2-2}P[S_n \ge \epsilon(n\log n)^{1/2}] < \infty \quad \textit{for all } \epsilon > \sigma(p-2)^{1/2},$$

$$(4.16)\quad \sum n^{p/2-2} P\left[\sup_{k\geq n}(k \log k)^{-1/2} S_k \geq \epsilon\right] < \infty \quad \textit{for all } \epsilon > \sigma(p-2)^{1/2},$$

$$(4.17)\quad \sum n^{p/2-2} P[S_n \geq \epsilon (n \log n)^{1/2}] < \infty \quad \textit{for some } \epsilon > 0.$$

5. Applications to the moments of the largest excess and the last time of boundary crossings. In this section, we shall consider moments of the last time $T(\epsilon, \alpha)$ and of the largest excess $M(\epsilon, \alpha)$ of certain boundary crossings for the sample sums S_n as defined in §1. These are related to the moments of $T_1(\epsilon, \alpha)$, $M_1(\epsilon, \alpha)$ for the original sample observations X_n. We now introduce the following notation: For $p > 0$, $\alpha > 0$,

$$(5.1)\quad J(\epsilon; p, \alpha) = \int_0^\infty t^{p\alpha-2} P[\overline{S}_t \geq \epsilon t^\alpha]\, dt,$$

$$(5.2)\quad I(\epsilon; p, \alpha) = \int_0^\infty t^{p\alpha-2} P\left[\sup_{k\geq t} k^{-\alpha} S_k \geq \epsilon\right] dt,$$

$$(5.3)\quad m(\epsilon; p, \alpha) = E(M(\epsilon, \alpha))^{(p\alpha-1)/\alpha}, \qquad M(\epsilon, \alpha) = \sup_{n\geq 0}(S_n - \epsilon n^\alpha),$$

$$(5.4)\quad r(\epsilon; p, \alpha) = E(T(\epsilon, \alpha))^{p\alpha-1}, \qquad T(\epsilon, \alpha) = \sup\{n \geq 1 : S_n \geq \epsilon n^\alpha\}\ (\sup \phi = 0),$$

$$(5.5)\quad s(\epsilon; p, \alpha) = E(\overline{S}_{T(\epsilon,\alpha)})^{(p\alpha-1)/\alpha},$$

$$(5.6)\quad J_1(\epsilon; p, \alpha) = \int_0^\infty t^{p\alpha-2} P[\overline{X}_t \geq \epsilon t^\alpha]\, dt,$$

$$(5.7)\quad I_1(\epsilon; p, \alpha) = \int_0^\infty t^{p\alpha-2} P\left[\sup_{k\geq t} k^{-\alpha} X_k \geq \epsilon\right] dt,$$

$$(5.8)\quad m_1(\epsilon; p, \alpha) = E(M_1(\epsilon, \alpha))^{(p\alpha-1)/\alpha}, \qquad M_1(\epsilon, \alpha) = \sup_{n\geq 0}(X_n - \epsilon n^\alpha),$$

$$(5.9)\quad r_1(\epsilon; p, \alpha) = E(T_1(\epsilon, \alpha))^{p\alpha-1}, \qquad T_1(\epsilon, \alpha) = \sup\{n \geq 1 : X_n \geq \epsilon n^\alpha\},$$

$$(5.10)\quad s_1(\epsilon; p, \alpha) = E(\overline{X}_{T_1(\epsilon,\alpha)})^{(p\alpha-1)/\alpha}.$$

Lemma 2. *Let* $S_1, S_2, \ldots$ *be any sequence of random variables (not necessarily sample sums). Define* $S_0 = \overline{S}_0 = 0$, $\overline{S}_t = \overline{S}_{[t]} = \max_{1\leq j\leq [t]} S_j$, *and define* $J(\epsilon; p, \alpha)$, $I(\epsilon; p, \alpha)$, $m(\epsilon; p, \alpha)$, $r(\epsilon; p, \alpha)$ *and* $s(\epsilon; p, \alpha)$ *as in* (5.1)–(5.5). *Then, for any positive constants* ϵ, p, α *with* $p\alpha > 1$,

$$(5.11)\quad \begin{aligned}(p\alpha - 1)\epsilon^{(p\alpha-1)/\alpha} I(2\epsilon; p, \alpha) &\leq m(\epsilon; p, \alpha)\\ &\leq (2^{(p\alpha-1)/\alpha} - 1)^{-1}(p\alpha - 1)\epsilon^{(p\alpha-1)/\alpha} J(\epsilon/2; p, \alpha),\end{aligned}$$

$$(5.12)\quad r(\epsilon; p, \alpha) \leq (p\alpha - 1) I(\epsilon; p, \alpha),$$

(5.13) $\epsilon^{(p\alpha-1)/\alpha}r(\epsilon; p, \alpha) \le s(\epsilon; p, \alpha) \le K_{p,\alpha}\{\epsilon^{(p\alpha-1)/\alpha}r(\epsilon; p, \alpha) + m(\epsilon; p, \alpha)\}$,

where $K_{p,\alpha} = 1$ *if* $p\alpha - 1 \le \alpha$ *and* $K_{p,\alpha} = 2^{((p\alpha-1)/\alpha)-1}$ *if* $p\alpha - 1 > \alpha$.

Proof. Let $r = (p\alpha - 1)/\alpha$. We note that

$$\epsilon^{-r}m(\epsilon; p, \alpha) = \int_0^\infty P\left[\sup_{n\ge 1}(S_n - \epsilon n^\alpha) \ge \epsilon t^{1/r}\right] dt$$

$$\le \sum_{k=1}^\infty \int_0^\infty P\left[\sup_{(2^{k-1}-1)t^{1/r} < n^\alpha \le (2^k-1)t^{1/r}} (S_n - \epsilon n^\alpha) \ge \epsilon t^{1/r}\right] dt$$

$$\le \sum_{k=1}^\infty \int_0^\infty P[\overline{S}_{(2^k t^{1/r})^{1/\alpha}} \ge 2^{k-1}\epsilon t^{1/r}]\, dt$$

$$= \sum_{k=1}^\infty 2^{-kr}(p\alpha - 1) \int_0^\infty u^{p\alpha-2}P[\overline{S}_u \ge (\epsilon/2)u^\alpha]\, du.$$

To complete the proof of (5.11), we have

$$I(2\epsilon; p, \alpha) = \int_0^\infty t^{p\alpha-2}P\left[\sup_{k\ge t} k^{-\alpha}S_k \ge 2\epsilon\right] dt$$

$$\le \int_0^\infty t^{p\alpha-2}P\left[\sup_{n\ge 0}(S_n - \epsilon n^\alpha) \ge \epsilon t^\alpha\right] dt = (p\alpha - 1)^{-1}E(M(\epsilon, \alpha)/\epsilon)^r.$$

Since $P[T(\epsilon, \alpha) \ge t] \le P[\sup_{k\ge t} k^{-\alpha}S_k \ge \epsilon]$, it is easy to see (5.12). Finally, (5.13) follows immediately from the fact that

$$\epsilon T^\alpha(\epsilon, \alpha) \le \overline{S}_{T(\epsilon,\alpha)} \le \epsilon T^\alpha(\epsilon, \alpha) + \sup_{n\ge 0}(S_n - \epsilon n^\alpha). \tag{5.14}$$

We remark that if $X_1, X_2, \ldots$ are i.i.d. random variables with $EX_1 = 0$ and (S_n) is the sequence of partial sums, then Lemma 2 and inequality (1.9) imply that for $\alpha > 1/2$, $p > 1/2$ and $\epsilon > 0$,

(5.15) $m(\epsilon; p, \alpha) \le A_{p,\alpha}\epsilon^{(p\alpha-1)/\alpha}\{E(X^+/\epsilon)^p + (E(X_1/\epsilon)^2)^{(p\alpha-1)/(2\alpha-1)}\}$

where $A_{p,\alpha} > 0$ is a universal constant depending only on p and α. In the case $p = 2$ and $\alpha = 1$, this reduces to $E(\sup_{n\ge 0}(S_n - \epsilon n)) \le A\epsilon^{-1}EX_1^2$, a result obtained by Kingman [10] by a different approach. In fact, Kingman showed that the constant A in the above upper bound can be taken to be 1/2. This is a very sharp bound for small ϵ in view of the fact that

$$\lim_{\epsilon \downarrow 0} \epsilon E\left(\sup_{n\ge 0}(S_n - \epsilon n)\right) = \tfrac{1}{2}EX_1^2$$

(see Theorem 7 below).

Lemma 3. *Suppose $\alpha > 1/2$ and $X_1, X_2, \ldots$ are i.i.d. random variables with $E|X_1|^{1/\alpha} < \infty$. Assume further that $EX_1 = 0$ in the case $\alpha \le 1$. Letting $S_n = X_1 + \cdots + X_n$, we have for any $\gamma > 0, \epsilon > 0$,*

$$E(T(\epsilon, \alpha))^\gamma < \infty \Rightarrow E(T_1(2\epsilon, \alpha))^\gamma < \infty.$$

Proof. Set $A_k = [X_k \ge 2\epsilon k^\alpha]$, $B_k = [|S_{k-1}| \le \epsilon k^\alpha]$. By the Marcinkiewicz-Zygmund strong law of large numbers, $n^{-\alpha}S_n \to 0$ a.e., and so $\lim_{k\to\infty} P(B_k) = 1$. Since $E|X_1|^{1/\alpha} < \infty$, $\lim_{m\to\infty} P(\bigcup_{j=m}^\infty A_j) = P(A_n \text{ i.o.}) = 0$, and so we can choose m_0 such that $P(B_k) - P(\bigcup_{j=m}^\infty A_j) \ge \frac{1}{2}$ if $k \ge m \ge m_0$. By an argument due to Erdös [5], we then obtain that, for $m \ge m_0$,

$$P[T(\epsilon, \alpha) \ge m] \ge \sum_{k=m}^\infty \left\{P(A_k \cap B_k) - P\left(A_k \cap \left(\bigcup_{j=m}^{k-1} A_j\right)\right)\right\}$$

$$\ge \frac{1}{2} \sum_{k=m}^\infty P(A_k) \ge \frac{1}{2} P[T_1(2\epsilon, \alpha) \ge m].$$

The desired conclusion then follows.

Lemma 4. *Suppose $X_1, X_2, \ldots$ are i.i.d. random variables and $\alpha > 0$, $p > 1/\alpha$. Then for all $\epsilon > 0$,*

$$(p\alpha - 1)J_1(\epsilon; p, \alpha) \le r_1(2^{-\alpha}\epsilon; p, \alpha) \quad \textit{and} \quad p\alpha J_1(\epsilon; p, \alpha) \le \epsilon^{-p}E(X_1^+)^p.$$

Furthermore, if $J_1(\epsilon; p, \alpha) < \infty$ for some $\epsilon > 0$, then $E(X_1^+)^p < \infty$ and

$$J_1(1; p, \alpha) \ge (E(X_1^+)^p - 1)/\{2p\alpha(1 + E(X_1^+)^{1/\alpha})\}. \tag{5.16}$$

Proof. We note that

$$r_1(2^{-\alpha}\epsilon; p, \alpha) = (p\alpha - 1) \int_0^\infty t^{p\alpha-2} P[T_1(2^{-\alpha}\epsilon, \alpha) \ge t]\, dt$$

$$= (p\alpha - 1) \int_0^\infty t^{p\alpha-2} P[X_n \ge 2^{-\alpha}\epsilon n^\alpha \text{ for some } n \ge t]\, dt$$

$$\ge (p\alpha - 1) \int_0^\infty t^{p\alpha-2} P\left[\max_{t<n\le 2t} X_n \ge \epsilon t^\alpha\right] dt$$

$$\ge (p\alpha - 1) \int_0^\infty t^{p\alpha-2} P[\overline{X}_t \ge \epsilon t^\alpha]\, dt.$$

It is also easy to see that

$$p\alpha J_1(\epsilon; p, \alpha) = p\alpha \int_1^\infty t^{p\alpha-2} P[\overline{X}_t \ge \epsilon t^\alpha]\, dt$$

$$\le p\alpha \int_1^\infty t^{p\alpha-1} P[X_1 \ge \epsilon t^\alpha]\, dt \le \epsilon^{-p} E(X_1^+)^p.$$

To prove (5.16), for any $c > 0$, define $X_i(c) = X_i I_{[X_i \le c]}$, $\overline{X}_t(c) = \max_{1\le i\le t} X_i(c)$, $\overline{X}_0(c) = X_0(c) = 0$. Put $n = [t]$, and note that for $t \ge 1$,

$$\begin{aligned}
P[\overline{X}_t(c) \ge t^\alpha] &= P[X_1(c) \ge t^\alpha] + P[X_2(c) \ge t^\alpha]P[X_1(c) < t^\alpha] \\
&\quad + \cdots + P[X_n(c) \ge t^\alpha]P[X_{n-1}(c) < t^\alpha] \cdots P[X_1(c) < t^\alpha] \\
&\ge nP[X_1(c) \ge t^\alpha]P^n[X_1(c) < t^\alpha] \\
&= nP[X_1(c) \ge t^\alpha](1 - P[\overline{X}_t \ge t^\alpha]).
\end{aligned}$$

Therefore, for $t \ge 1$,

$$\begin{aligned}
\tfrac{1}{2} tP[X_1(c) \ge t^\alpha] &\le nP[X_1(c) \ge t^\alpha] \le P[\overline{X}_t(c) \ge t^\alpha]\{1 + nP[X_1(c) \ge t^\alpha]\} \\
&\le P[\overline{X}_t(c) \ge t^\alpha]\{1 + E(X_1^+(c))^{1/\alpha}\}, \quad \text{by the Markov inequality.}
\end{aligned}$$

From this, it follows that

$$\begin{aligned}
E(X_1^+(c))^p - 1 &\le p\alpha \int_1^\infty t^{p\alpha-1} P[X_1(c) \ge t^\alpha]\,dt \\
&\le 2p\alpha\{1 + E(X_1^+(c))^{1/\alpha}\} \int_1^\infty t^{p\alpha-2} P[\overline{X}_t(c) \ge t^\alpha]\,dt.
\end{aligned} \tag{5.17}$$

Suppose $E(X_1^+)^p = \infty$. Then $E(X_1^+(c))^{1/\alpha} = o(E(X_1^+(c))^p)$ as $c \to \infty$, and so it easily follows from (5.17) that $J_1(1; p, \alpha) = \infty$. Hence $J_1(1; p, \alpha) < \infty$ implies that $E(X_1^+)^p < \infty$, and in this case, letting $c \uparrow \infty$ in (5.17), we obtain (5.16).

Lemmas 2, 3 and 4 together with the results in §3 give the following theorem.

Theorem 5. *Suppose $X_1, X_2, \ldots$ are i.i.d. random variables, $S_n = X_1 + \cdots + X_n$, and $\alpha > 0$, $p > 0$ such that $\alpha p > 1$.*

(i) *If $E(X_1^+)^p < \infty$, then for every $\epsilon > 0$, $I_1(\epsilon; p, \alpha)$, $J_1(\epsilon; p, \alpha)$, $m_1(\epsilon; p, \alpha)$, $r_1(\epsilon; p, \alpha)$ and $s_1(\epsilon; p, \alpha)$ are all finite. Conversely, if one of the above five quantities is finite for some $\epsilon > 0$, then $E(X_1^+)^p < \infty$.*

(ii) *Suppose $E(X_1^+)^p < \infty$ and $\alpha > \frac{1}{2}$. In the case $\alpha = 1$, assume further that $EX_1 = 0$. In the case $\alpha < 1$, assume further that $EX_1 = 0$ and $E|X_1|^r < \infty$ for some $2 \ge r > 1/\alpha$. Under these assumptions, $I(\epsilon; p, \alpha)$, $J(\epsilon; p, \alpha)$, $m(\epsilon; p, \alpha)$, $r(\epsilon; p, \alpha)$ and $s(\epsilon; p, \alpha)$ are finite for all $\epsilon > 0$.*

(iii) *Suppose $\alpha > \frac{1}{2}$ and $E|X_1|^{1/\alpha} < \infty$. Assume further that $EX_1 = 0$ in the case $\alpha \le 1$. If one of $I(\epsilon; p, \alpha)$, $J(\epsilon; p, \alpha)$, $m(\epsilon; p, \alpha)$, $r(\epsilon; p, \alpha)$ and $s(\epsilon; p, \alpha)$ is finite for some $\epsilon > 0$, then $E(X_1^+)^p < \infty$.*

6. The limiting distribution and limiting moments of the last time and largest excess of boundary crossings for sample sums. The following theorem gives the limiting distribution of $T(\epsilon,\alpha)$, $M(\epsilon,\alpha)$ and $S_{T(\epsilon,\alpha)}$ as $\epsilon \downarrow 0$.

Theorem 6. *Suppose $W(t)$, $t \geq 0$, is the standard Wiener process and X_1, $X_2, \ldots$ are i.i.d. random variables with $EX_1 = 0$, $EX_1^2 = 1$. Let $S_n = X_1 + \cdots + X_n$, $\alpha > \frac{1}{2}$, and define $M(\epsilon,\alpha)$ and $T(\epsilon,\alpha)$ for any $\epsilon > 0$ by (5.3) and (5.4). Let $M^*(\alpha) = \sup_{t\geq 0}(W(t) - t^\alpha)$, $T^*(\alpha) = \sup\{t \geq 0 : W(t) \geq t^\alpha\}$. Then as $\epsilon \downarrow 0$,*

$$\epsilon^{2/(2\alpha-1)} T(\epsilon,\alpha) \xrightarrow{\mathfrak{D}} T^*(\alpha), \tag{6.1}$$

$$\epsilon^{1/(2\alpha-1)} M(\epsilon,\alpha) \xrightarrow{\mathfrak{D}} M^*(\alpha), \tag{6.2}$$

$$\epsilon^{1/(2\alpha-1)} S_{T(\epsilon,\alpha)} \xrightarrow{\mathfrak{D}} (T^*(\alpha))^\alpha. \tag{6.3}$$

Proof. (6.1) and (6.2) can be proved by an argument similar to that of Robbins, Siegmund and Wendel [15] or that of Müller, [19], who considered the problem in the case $\alpha = 1$. Alternatively we can use a result of Robbins and Siegmund [14, Theorem 2] in the following way. For any $x > 0$,

$$\begin{aligned} &P[\epsilon^{2/(2\alpha-1)} T(\epsilon,\alpha) \geq x] \\ &\quad = P[S_n \geq \epsilon n^\alpha \text{ for some } n \geq \epsilon^{-2/(2\alpha-1)} x] \\ &\quad = P[S_n \geq \sqrt{m}(n/m)^\alpha \text{ for some } n \geq xm], \quad \text{where } m = \epsilon^{-2/(2\alpha-1)}, \\ &\quad \to P[W(t) \geq t^\alpha \text{ for some } t \geq x] \quad \text{as } m \to \infty. \end{aligned}$$

Similarly, given any $x > 0$, if we apply part (ii) of Theorem 2 in [14], where we set $g(t) = t^\alpha + x$, then

$$\begin{aligned} P[\epsilon^{1/(2\alpha-1)} M(\epsilon,\alpha) > x] &= P\left[m^{-1/2} \sup_{n\geq 1}(S_n - \sqrt{m}(n/m)^\alpha) > x\right], \\ &\qquad\qquad \text{where } m = \epsilon^{-2/(2\alpha-1)}, \\ &= P[m^{-1/2} S_n > (n/m)^\alpha + x \text{ for some } n \geq 1] \\ &\to P[W(t) > t^\alpha + x \text{ for some } t > 0] \quad \text{as } m \to \infty. \end{aligned}$$

To prove (6.3), we note that $S_{T(\epsilon,\alpha)} \leq \epsilon(T(\epsilon,\alpha) + 1)^\alpha - X_{T(\epsilon,\alpha)+1}$ and so

$$\begin{aligned} &\{\epsilon^{2/(2\alpha-1)}(T(\epsilon,\alpha) + 1)\}^\alpha - \epsilon^{1/(2\alpha-1)} X_{T(\epsilon,\alpha)+1} \\ &\qquad \geq \epsilon^{1/(2\alpha-1)} S_{T(\epsilon,\alpha)} \geq (\epsilon^{2/(2\alpha-1)} T(\epsilon,\alpha))^\alpha. \end{aligned} \tag{6.4}$$

Since $EX_1^2 < \infty$ implies that $n^{-1/2}X_n \to 0$ a.e., it follows that

$$(T(\epsilon, \alpha))^{-1/2} X_{T(\epsilon,\alpha)+1} \xrightarrow{P} 0.$$

By (6.1), $\epsilon^{1/(2\alpha-1)}(T(\epsilon, \alpha))^{1/2} \xrightarrow{\mathfrak{D}} (T^*(\alpha))^{1/2}$. Therefore

$$\epsilon^{1/(2\alpha-1)} X_{T(\epsilon,\alpha)+1} \xrightarrow{P} 0. \tag{6.5}$$

From (6.1), (6.4) and (6.5), it is easy to see that (6.3) holds.

By making use of the inequality (1.9), we now obtain from Theorem 6 the limiting moments of $T(\epsilon, \alpha)$, $M(\epsilon, \alpha)$ and $S_{T(\epsilon,\alpha)}$ as $\epsilon \downarrow 0$.

Theorem 7. *With the same notations and assumptions as in Theorem* 6, *if* $E(X_1^+)^p < \infty$ *for some* $p > 2$, *then for any* $\alpha > \frac{1}{2}$,

$$\lim_{\epsilon \downarrow 0} \epsilon^{2(p\alpha-1)/(2\alpha-1)} E(T(\epsilon, \alpha))^{p\alpha-1} = E(T^*(\alpha))^{p\alpha-1},$$

$$\lim_{\epsilon \downarrow 0} \epsilon^{(p\alpha-1)/\{\alpha(2\alpha-1)\}} E(M(\epsilon, \alpha))^{(p\alpha-1)/\alpha} = E(M^*(\alpha))^{(p\alpha-1)/\alpha},$$

$$\lim_{\epsilon \downarrow 0} \epsilon^{(p\alpha-1)/\{\alpha(2\alpha-1)\}} E(S_{T(\epsilon,\alpha)})^{(p\alpha-1)/\alpha} = E(T^*(\alpha))^{p\alpha-1}.$$

The above relations also hold for $2 \geq p > 1/\alpha$.

Proof. In view of Theorem 6, we need only show that $(\epsilon^{2/(2\alpha-1)}T(\epsilon, \alpha))^{p\alpha-1}$ and $(\epsilon^{1/(2\alpha-1)}M(\epsilon, \alpha))^{(p\alpha-1)/\alpha}$, $0 < \epsilon \leq 1$, are uniformly integrable, as this will in turn imply that $(\epsilon^{1/(2\alpha-1)} S_{T(\epsilon,\alpha)})^{(p\alpha-1)/\alpha}$, $0 < \epsilon \leq 1$, is also uniformly integrable by the inequality (5.14). First consider the case $p \geq 2$. Let $0 < \delta < \frac{1}{2}$ and define

$$X_i' = X_i I_{[|X_i| \leq K]} - EX_i I_{[|X_i| \leq K]}, \qquad X_i'' = X_i I_{[|X_i| > K]} - EX_i I_{[|X_i| > K]}, \tag{6.6}$$

where $K > 0$ is so chosen that

$$E((X_1''/\delta)^+)^p + (E(X_1''/\delta)^2)^{(p\alpha-1)/(2\alpha-1)} \leq \delta. \tag{6.7}$$

Let

$$S_n' = X_1' + \cdots + X_n', \; S_n'' = X_1'' + \cdots + X_n''$$

and define $T'(\epsilon, \alpha)$, $M'(\epsilon, \alpha)$ for S_n', $T''(\epsilon, \alpha)$ and $M''(\epsilon, \alpha)$ for S_n''. Using Lemma 2 and (1.9), we obtain that for $0 < \epsilon \leq 1$,

$$\begin{aligned}
&\epsilon^{2(p\alpha-1)/(2\alpha-1)}E(T''(\delta\epsilon,\alpha))^{p\alpha-1}\\
&\quad\le A_{p,\alpha}\epsilon^{2(p\alpha-1)/(2\alpha-1)}\sum_{n=1}^{\infty} n^{p\alpha-2}P\Big[\max_{1\le j\le n} S_j'' \ge (\delta\epsilon/4)n^{\alpha}\Big]\\
(6.8)\qquad &\quad\le B_{p,\alpha}\epsilon^{2(p\alpha-1)/(2\alpha-1)}\{E((X_1''/\delta\epsilon)^+)^p + (E(X_1''/\delta\epsilon)^2)^{(p\alpha-1)/(2\alpha-1)}\}\\
&\quad\le B_{p,\alpha}\delta, \quad \text{noting that } p\le 2(p\alpha-1)/(2\alpha-1) \text{ since } p\ge 2.
\end{aligned}$$

The constants $A_{p,\alpha}$ and $B_{p,\alpha}$ above depend only on p and α. Now take any $q > p$ and by a similar argument as in (6.8), we have

$$\begin{aligned}
&E(\epsilon^{2/(2\alpha-1)}T'((1-\delta)\epsilon,\alpha))^{q\alpha-1}\\
(6.9)\qquad &\quad\le C_{q,\alpha}\{E((X_1')^+)^q + (E(X_1')^2)^{(q\alpha-1)/(2\alpha-1)}\},
\end{aligned}$$

where $C_{q,\alpha}$ is a positive constant depending only on q and α. Hence setting $Z(\epsilon,\delta) = (\epsilon^{2/(2\alpha-1)}T'((1-\delta)\epsilon,\alpha))^{p\alpha-1}$, we have the uniform integrability of $Z(\epsilon,\delta)$, $0<\epsilon\le 1$, and so we can choose $\eta > 0$ such that if $P(A) < \eta$, then $EZ(\epsilon,\delta)I_A < \delta$ for all $0<\epsilon\le 1$. Since $T(\epsilon,\alpha)\le T'((1-\delta)\epsilon,\alpha) + T''(\delta\epsilon,\alpha)$, we have established the uniform integrability of $(\epsilon^{2/(2\alpha-1)}T(\epsilon,\alpha))^{p\alpha-1}$, $0<\epsilon\le 1$, in the case $p\ge 2$.

Now let $2 > p > 1/\alpha$. Then by what we just proved, $(\epsilon^{2/(2\alpha-1)}T(\epsilon,\alpha))^{2\alpha-1}$ is uniformly integrable, and so $(\epsilon^{2/(2\alpha-1)}T(\epsilon,\alpha))^{p\alpha-1}$ is also uniformly integrable. The desired conclusion for $M(\epsilon,\alpha)$ can be similarly proved.

Theorem 7 gives the asymptotic behavior of $r(\epsilon;p,\alpha)$ and $m(\epsilon;p,\alpha)$ as $\epsilon\downarrow 0$. It is also interesting to investigate the asymptotic behavior of $J(\epsilon;p,\alpha)$ and $I(\epsilon;p,\alpha)$. This is given in the following theorem.

Theorem 8. *With the same notations and assumptions as in Theorem 6, define $J(\epsilon;p,\alpha)$, $I(\epsilon;p,\alpha)$ by (5.1) and (5.2), and let Φ denote the distribution function of the standard normal distribution. If $E(X_1^+)^p<\infty$ for some $p\ge 2$, then for any $\alpha > ½$,*

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon^{2(p\alpha-1)/(2\alpha-1)}J(\epsilon;p,\alpha) = 2\int_0^\infty t^{p\alpha-2}(1-\Phi(t^{(2\alpha-1)/2}))\,dt\\
(6.10)\qquad &\quad= 2\lim_{\epsilon\downarrow 0}\epsilon^{2(p\alpha-1)/(2\alpha-1)}\int_0^\infty t^{p\alpha-2}P[S_t\ge\epsilon t^{\alpha}]\,dt,
\end{aligned}$$

$$(6.11)\qquad \lim_{\epsilon\downarrow 0}(p\alpha-1)\epsilon^{2(p\alpha-1)/(2\alpha-1)}I(\epsilon;p,\alpha) = E(T^*(\alpha))^{p\alpha-1}.$$

Proof. To prove (6.10), we note that by a change of variable, we can write

$$J(\epsilon; p, \alpha)$$

$$= \epsilon^{-2(p\alpha-1)/(2\alpha-1)} \tag{6.12}$$

$$\cdot \int_0^\infty u^{p\alpha-2} P[\overline{S}_{u\epsilon^{-2/(2\alpha-1)}} \geq \epsilon^{-1/(2\alpha-1)} u^\alpha]\, du.$$

By Donsker's invariance principle, for any $u > 0$,

$$\lim_{\epsilon \downarrow 0} P[\overline{S}_{u\epsilon^{-2/(2\alpha-1)}} \geq \epsilon^{-1/(2\alpha-1)} u^\alpha]$$

$$= P\left[\max_{0 \leq t \leq 1} W(t) \geq u^{\alpha-\frac{1}{2}}\right] = 2(1 - \Phi(u^{\alpha-\frac{1}{2}})). \tag{6.13}$$

It then follows from (6.12), (6.13) and Fatou's lemma that

$$\liminf_{\epsilon \downarrow 0} \epsilon^{2(p\alpha-1)/(2\alpha-1)} J(\epsilon; p, \alpha) \tag{6.14}$$

$$\geq 2 \int_0^\infty u^{p\alpha-2}(1 - \Phi(u^{(2\alpha-1)/2}))\, du.$$

To obtain the reverse inequality with lim inf replaced by lim sup, we let $0 < \delta < 1$ and define X_i', X_i'' by (6.6) with $K > 0$ so chosen that (6.7) is satisfied and $0 < \sigma \leq 1 + \delta$, where $\sigma^2 = E(X_1')^2$. Let $S_n' = X_1' + \cdots + X_n'$, $S_n'' = X_1'' + \cdots + X_n''$ and define $J'(\epsilon; p, \alpha)$ for S_n', $J''(\epsilon; p, \alpha)$ for S_n''. Obviously $J(\epsilon; p, \alpha) \leq J'((1-\delta)\epsilon; p, \alpha) + J''(\delta\epsilon; p, \alpha)$. Using the inequality (1.9) as in (6.8), we obtain that

$$\epsilon^{2(p\alpha-1)/(2\alpha-1)} J''(\delta\epsilon; p, \alpha)$$

$$\leq \xi_{p,\alpha} \epsilon^{2(p\alpha-1)/(2\alpha-1)} \sum_{n=1}^\infty n^{p\alpha-2} P\left[\max_{1 \leq j \leq n} (S_j''/(\delta\epsilon)) \geq n^\alpha\right] \tag{6.15}$$

$$\leq \zeta_{p,\alpha}\delta, \quad \text{where } \xi_{p,\alpha} \text{ and } \zeta_{p,\alpha} \text{ depend only on } p \text{ and } \alpha.$$

Choose $B \geq 1$ large enough that $u^\alpha - (2u)^{\frac{1}{2}} > \frac{1}{2}u^\alpha$ for all $u \geq B$ and

$$2 \int_B^\infty u^{p\alpha-2}(1 - \Phi(\tfrac{1}{2}u^{\alpha-\frac{1}{2}}))\, du \leq \delta.$$

Let $\tilde{\epsilon} = (1-\delta)\epsilon/\sigma$. As in (6.12), we have

$$\tilde{\epsilon}^{2(p\alpha-1)/(2\alpha-1)} J'((1-\delta)\epsilon; p, \alpha)$$

$$= \int_0^\infty u^{p\alpha-2} P\left[\max_{j \leq u\tilde{\epsilon}^{-2/(2\alpha-1)}} (S_j'/\sigma) \geq \tilde{\epsilon}^{-1/(2\alpha-1)} u^\alpha\right] du. \tag{6.16}$$

By the dominated convergence theorem,

$$
\begin{aligned}
&\int_0^B u^{p\alpha-2} P\left[\max_{j\le u\tilde{\epsilon}^{-2/(2\alpha-1)}} (S_j'/\sigma) \ge \tilde{\epsilon}^{-1/(2\alpha-1)} u^{\alpha}\right] du \\
(6.17)\qquad &\to \int_0^B u^{p\alpha-2} P\left[\max_{0\le t\le 1} W(t) \ge u^{\alpha-1/2}\right] du \quad (\text{as } \epsilon \to 0) \\
&\le 2\int_0^{\infty} u^{p\alpha-2}(1-\Phi(u^{(2\alpha-1)/2}))\,du.
\end{aligned}
$$

Using the Lévy inequality [12, p. 248], we obtain for $u \ge B$,

$$
\begin{aligned}
&P\left[\max_{j\le u\tilde{\epsilon}^{-2/(2\alpha-1)}} (S_j'/\sigma) \ge \tilde{\epsilon}^{-1/(2\alpha-1)} u^{\alpha}\right] \\
(6.18)\qquad &\le 2P[S'_{[u\tilde{\epsilon}^{-2/(2\alpha-1)}]}/\sigma \ge \tfrac{1}{2}\tilde{\epsilon}^{-1/(2\alpha-1)}u^{\alpha}], \quad \text{since } u^{\alpha} - (2u)^{1/2} > \tfrac{1}{2}u^{\alpha}.
\end{aligned}
$$

We now apply an estimate, due to Esseen [6], to the above probability: Let k be a positive integer such that $k-2 > 2(p\alpha-1)$. Then there exist positive constants c_1, c_2 depending on the absolute moments $E|X_1'|^2, E|X_1'|^3, \cdots, E|X_1'|^k$ such that for all $n = 1, 2, \cdots$ and all real x,

$$|P[S_n' < \sigma\sqrt{n}x] - \Phi(x)| \le c_1 n^{-1/2}(1+|x|^3)e^{-x^2/2} + c_2 n^{-(k-2)/2}$$

(cf. [6, pp. 73–76]). Set $n = [u\tilde{\epsilon}^{-2/(2\alpha-1)}]$ for $u \ge B$. Then $n \ge 1$ if $\tilde{\epsilon} < 1$. Therefore applying the above estimate, we have for $\tilde{\epsilon} < 1$ and $u \ge B$,

$$
\begin{aligned}
&P[S_n'/\sigma \ge \tfrac{1}{2}\tilde{\epsilon}^{-1/(2\alpha-1)}u^{\alpha}] \le P[S_n' \ge \tfrac{1}{2}\sigma\sqrt{n}u^{\alpha-1/2}] \\
(6.19)\qquad &\le 1 - \Phi(\tfrac{1}{2}u^{\alpha-1/2}) + c_1 n^{-1/2}(1+u^{3(\alpha-1/2)})\exp(-(1/8)u^{2\alpha-1}) \\
&\quad + 2^{(k-2)/2}c_2\tilde{\epsilon}^{(k-2)/(2\alpha-1)}u^{-(k-2)/2}.
\end{aligned}
$$

Since $(k-2)/2 > p\alpha - 1$, it then follows from (6.19) that

$$
\begin{aligned}
&\limsup_{\epsilon\downarrow 0} 2\int_B^{\infty} P[S'_{[u\tilde{\epsilon}^{-2/(2\alpha-1)}]}/\sigma \ge \tfrac{1}{2}\tilde{\epsilon}^{-1/(2\alpha-1)}u^{\alpha}]u^{p\alpha-2}\,du \\
(6.20)\qquad &\le 2\int_B^{\infty}(1-\Phi(\tfrac{1}{2}u^{\alpha-1/2}))u^{p\alpha-2}\,du \le \delta.
\end{aligned}
$$

It then follows from (6.15), (6.16), (6.17), (6.18) and (6.20) that

$$\limsup_{\epsilon\downarrow 0} \epsilon^{2(p\alpha-1)/(2\alpha-1)} J(\epsilon;\, p, \alpha)$$

$$\leq \zeta_{p,\alpha}\delta + (\sigma/(1-\delta))^{2(p\alpha-1)/(2\alpha-1)}\left\{2\int_0^\infty u^{p\alpha-2}(1-\Phi(u^{(2\alpha-1)/2}))\,du+\delta\right\}.$$

Since $\sigma \leq 1+\delta$ and δ is arbitrary, we therefore have

$$\limsup_{\epsilon\downarrow 0} \epsilon^{2(p\alpha-1)/(2\alpha-1)} J(\epsilon;\, p, \alpha) \leq 2\int_0^\infty u^{p\alpha-2}(1-\Phi(u^{(2\alpha-1)/2}))\,du.$$

In a similar way, we can obtain the asymptotic behavior of

$$\int_0^\infty t^{p\alpha-2} P[S_t \geq \epsilon t^{\alpha}]\,dt$$

as $\epsilon \downarrow 0$ given in (6.10). Finally, we note that

$$P\left[\sup_{k\geq t} k^{-\alpha} S_k \geq \epsilon\right] \geq P[T(\epsilon,\alpha) \geq t] \geq P\left[\sup_{k\geq t} k^{-\alpha} S_k > \epsilon\right]$$

and so (6.11) follows immediately from Theorem 7.

REFERENCES

1. L. E. Baum and M. Katz, *Convergence rates in the law of large numbers*, Trans. Amer. Math. Soc. 120 (1965), 108–123. MR 33 #6679.
2. Y. S. Chow, *Delayed sums and Borel summability for independent, identically distributed random variables*, Bull. Inst. Math. Acad. Sinica 1 (1973), 207–220.
3. Y. S. Chow, H. Robbins and D. Siegmund, *Great expectations: The theory of optimal stopping*, Houghton-Mifflin, Boston, Mass., 1971.
4. J. L. Doob, *Stochastic processes*, Wiley, New York; Chapman & Hall, London, 1953. MR 15, 445.
5. P. Erdös, *On a theorem of Hsu and Robbins*, Ann. Math. Statist. 20 (1949), 286–291. MR 11, 40.
6. C. -G. Esseen, *Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law*, Acta Math. 77 (1945), 1–125. MR 7, 312.
7. P. Hartman and A. Wintner, *On the law of the iterated logarithm*, Amer. J. Math. 63 (1941), 169–176. MR 2, 228.
8. C. S. Kao, *On the time and the excess of linear boundary crossings of the sample sums*, Ph. D. Thesis, Columbia University, New York, 1972.
9. J. Kiefer and J. Wolfowitz, *On the characteristics of the general queueing process, with applications to random walk*, Ann. Math. Statist. 27 (1956), 147–161. MR 17, 980.
10. J. F. C. Kingman, *Some inequalities for the queue GI/G/1*, Biometrika 49 (1962), 315–324. MR 33 #6720.
11. T. L. Lai, *Limit theorems for delayed sums*, Ann. Probability 2 (1974), 432–440.
12. M. Loève, *Probability theory. Foundations. Random sequences*, 3rd. ed., Van Nostrand, Princeton, N. J., 1963. MR 34 #3596.
13. J. Marcinkiewicz et A. Zygmund, *Quelques théorèmes sur les fonctions indépendantes*, Studia Math. 7 (1938), 104–120.
14. H. Robbins and D. Siegmund, *Boundary crossing probabilities for the Wiener process and sample sums*, Ann. Math. Statist. 41 (1970), 1410–1429. MR 43 #2796.

15. H. Robbins, D. Siegmund and J. Wendel, *The limiting distribution of the last time $S_n \geq n\epsilon$*, Proc. Nat. Acad. Sci. U.S.A. 61 (1968), 1228–1230. MR 39 #4946.

16. F. Spitzer, *A combinatorial lemma and its application to probability theory*, Trans. Amer. Math. Soc. 82 (1956), 323–339. MR 18, 156.

17. A. Zygmund, *Trigonometric series*. Vols. 1, 2, 2nd ed., Cambridge Univ. Press, New York, 1959. MR 21 #6498.

18. L. Baum, *On convergence to $+\infty$ in the law of large numbers*, Ann. Math. Statist. 34 (1963), 219–222. MR 26 #799.

19. D. W. Müller, *Verteilungs–Invarianzprinzipien für das starke Gesetz der grossen Zahl*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 10 (1968), 173–192. MR 38 #753.

DEPARTMENT OF MATHEMATICAL STATISTICS, COLUMBIA UNIVERSITY, NEW YORK, NEW YORK 10027

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

NECESSARY CONDITIONS FOR ISOMORPHISM OF LIE ALGEBRAS OF BLOCK

BY

JOHN B. JACOBS

ABSTRACT. Two algebras of Block, $\mathfrak{L}(G, \delta, f)$ and $\mathfrak{L}(G', \delta', f')$, are isomorphic only if $m(G) = m(G')$. This is not sufficient for isomorphism.

Let $\mathfrak{L}$ be a simple finite-dimensional Lie algebra over Φ, an algebraically closed field of prime characteristic p. Simplicity allows the identification $x \leftrightarrow \operatorname{ad} x$ for each $x \in \mathfrak{L}$. (That $\mathfrak{L}$ be centerless is sufficient for the identification.) Then if $\mathfrak{D}(\mathfrak{L})$ denotes the derivation algebra of $\mathfrak{L}$ we have $\mathfrak{L}$ $(= \operatorname{ad} \mathfrak{L}) \subset \mathfrak{D}(\mathfrak{L})$. For each $x \in \mathfrak{L}$, $(\operatorname{ad} x)^p$ is a derivation of $\mathfrak{L}$ and, if $(\operatorname{ad} \mathfrak{L})^{p^k}$ is the vector space spanned by $\{(\operatorname{ad} x)^{p^k} | x \in \mathfrak{L}\}$, then $\mathfrak{R}(\mathfrak{L}) = \operatorname{ad} \mathfrak{L} + (\operatorname{ad} \mathfrak{L})^p + (\operatorname{ad} \mathfrak{L})^{p^2} + \cdots$ is a subalgebra of $\mathfrak{D}(\mathfrak{L})$ which is restricted. We will call $\mathfrak{R}(\mathfrak{L})$ the restricted algebra of $\mathfrak{L}$. If $\mathfrak{L}$ is restricted, then $\operatorname{ad} \mathfrak{L} = \mathfrak{R}(\mathfrak{L})$, or under the identification, $\mathfrak{L} = \mathfrak{R}(\mathfrak{L})$. Thus, for any arbitrary centerless algebra $\mathfrak{L}$, $\mathfrak{L} \subseteq \mathfrak{R}(\mathfrak{L}) \subseteq \mathfrak{D}(\mathfrak{L})$. Clearly, any two isomorphic simple algebras $\mathfrak{L}$ and $\mathfrak{L}'$ over Φ must have $\mathfrak{R}(\mathfrak{L}) \cong \mathfrak{R}(\mathfrak{L}')$ and $\mathfrak{D}(\mathfrak{L}) \cong \mathfrak{D}(\mathfrak{L}')$. We will use this relationship to determine isomorphism conditions upon the algebras of Block.

Let G be an elementary abelian p-group written as a direct summand of elementary abelian p-groups, $G = G_0 \oplus G_1 \oplus \cdots \oplus G_m$. Let Φ be an algebraically closed field of characteristic $p > 3$. For each $i = 0, 1, \ldots, m$ define $f: G \times G \to \Phi$ such that $f|_{G_i} = f_i: G_i \times G_i \to \Phi$ is a skew-symmetric, nondegenerate biadditive form. Then $f = f_0 + f_1 + \cdots + f_m$. For each $i = 1, \ldots, m$, assume that there exist additive functions $g_i, h_i: G_i \to \Phi$ such that $f_i(\alpha, \beta) = g_i(\alpha)h_i(\beta) - g_i(\beta)h_i(\alpha)$. Pick $\delta_i \in G_i$ for which $g_i(\delta_i) = 0$, and set $\delta = \delta_1 + \cdots + \delta_m$. Define $\mathfrak{L}(G, \delta, f)$ to be the Lie algebra over Φ with basis $\{u_\alpha | \alpha \in G, \alpha \neq 0, -\delta\}$ where multiplication is given by

$$u_\alpha u_\beta = \sum_{i=0}^{m} f_i(\alpha_i, \beta_i) u_{\alpha+\beta-\delta_i}.$$

Received by the editors January 17, 1974.
AMS (MOS) subject classification (1970). Primary 17B40.
Key words and phrases. Algebras of Block, restricted algebra, isomorphism.

Here α_i and β_i denote the ith components of α and β, respectively, in G and δ_0 is assumed to be zero. $\mathfrak{L}(G, \delta, f)$ is then a simple algebra over Φ called an algebra of Block.

The derivations of the algebras of Block have been completely determined in [1]. As they will be utilized later, a brief description follows.

Since G is an elementary abelian p-group it is an n-dimensional vector space over Φ_p (the prime subfield of $\Phi = GF(p)$), each of the G_i's being a subspace of dimension, say, n_i. Pick a basis $\{\sigma_{01}, \sigma_{02}, \dots, \sigma_{0n_0}\}$ for G_0 and $\{\sigma_{i1}, \dots, \sigma_{in_i-1}, \delta_i\}$ for G_i, $i = 1, \dots, m$, such that $f(\sigma_{i1}, \delta_i) = f_i(\sigma_{i1}, \delta_i) \neq 0$. Such is possible since f is nondegenerate. For each $\alpha \in G$, write $\alpha = \Sigma_{i,j}\, s_{ij}(\alpha)\sigma_{ij} + \Sigma_i\, s_i(\alpha)\delta_i$. The coefficients $s_{ij}(\alpha)$, $s_i(\alpha)$ of the σ_{ij}'s and δ_i's are unique since the σ_{ij}'s and δ_i's form a basis of G. The derivations of $\mathfrak{L}(G, \delta, f)$ are linear combinations (over Φ) of the elements in the following sets:

(i) $R = \{\text{ad}\, u_\alpha \mid \alpha \in G,\ \alpha \neq 0\}$ (ad $u_{-\delta}$ is included although not an element of $\mathfrak{L}(G, \delta, f)$).

(ii) $S = \{D(\sigma_{k1}, -\delta_k), D(\delta_k, -\delta_k) \mid k = 1, \dots, m\}$ where $u_\alpha D(\gamma_k, -\delta_k) = f(\alpha, \gamma_k)u_{\alpha-\delta_k}$ for γ_k in G_k.

(iii) $T = \{D(\sigma_{0k}, 0), D(\sigma_{ij}, 0) \mid k = 1, \dots, n_0;\ i = 1, \dots, m;\ j = 2, \dots, n_i - 1\}$ where $G_0 \neq \{0\}$ and $T = \{D(\delta, 0), D(\sigma_{ij}, 0) \mid i = 1, \dots, m;\ j = 2, \dots, n_i - 1\}$ when $G_0 = \{0\}$; where $u_\alpha D(\sigma_{ij}, 0) = s_{ij}(\alpha)u_\alpha$ and

$$u_\alpha D(\delta, 0) = \left(-1 + \sum_i s_i(\alpha)\right)u_\alpha.$$

($D(\sigma_{i1}, 0)$ is a linear combination of ad u_{δ_i} and the remaining $D(\sigma_{ij}, 0)$'s.)

The set S is, of course, empty when $m = 0$. The dimension of $\mathfrak{L}(G,\delta,f)$ is $np^n - 1$ for $m = 0$ and $np^n - 2$ for $m > 0$, and it follows that the dimension of its derivation algebra, $\mathfrak{D}(\mathfrak{L}(G, \delta, f))$, is

(i) $np^n + n - 1$ when $G = G_0$ or when $G_0 \neq 0$ and $m > 0$.

(ii) $np^n + n$ when $G_0 = 0$.

From the dimensions of the derivation algebras and their derived algebras, Block concludes in [1, Theorem 14, Corollary 1] that necessary conditions for two algebras $\mathfrak{L}(G, \delta, f)$ and $\mathfrak{L}(G', \delta', f')$ to be isomorphic are that either $G_0 = 0$, $G'_0 = 0$, and $m(G) = m(G')$; or $G_0 \neq 0 \neq G'_0$ and $\min\{2, m(G)\} = \min\{2, m(G')\}$. By considering the restricted algebra of $\mathfrak{L}(G, \delta, f)$ we will show that it is necessary that $m(G) = m(G')$ for isomorphism and that, indeed, this is not sufficient.

For $u_\alpha, u_\beta \in \mathfrak{L}(G, \delta, f)$ it is easily shown by induction on m that

$$u_\alpha(\mathrm{ad}\,u_\beta)^p = \sum_{i=0}^m f(\alpha_i, \beta_i)f(\alpha_i - \delta_i, \beta_i)\cdots f(\alpha_i - (p-1)\delta_i, \beta_i)u_\alpha$$

The following lemma then shows that

$$u_\alpha(\mathrm{ad}\,u_\beta)^p = \sum_{i=0}^m \{f(\alpha_i, \beta_i)^p - f(\alpha_i, \beta_i)f(\delta_i, \beta_i)^{p-1}\}u_\alpha$$

Lemma 1. *Let* $a, b \in \Phi$, char $\Phi = p > 0$. *Then*

$$a(a-b)(a-2b)\cdots(a-(p-1)b) = a^p - ab^{p-1}.$$

Proof. The polynomial $x^p - xb^{p-1}$ has roots ib for $i = 0, \ldots, p-1$. Hence, $x^p - xb^{p-1} = \Pi_{i=0}^{p-1}(x - ib)$. Substituting a for x yields the desired result.

It is evident that

$$u_\alpha(\mathrm{ad}\,u_\beta)^{p^2} = \left\{\sum_{i=0}^m f(\alpha_i, \beta_i)^p - f(\alpha_i, \beta_i)f(\delta_i, \beta_i)^{p-1}\right\}^p u_\alpha$$

$$= \sum_{i=0}^m \{f(\alpha_i, \beta_i)^{p^2} - f(\alpha_i, \beta_i)^p f(\delta_i, \beta_i)^{p(p-1)}\}u_\alpha,$$

and more generally that

$$u_\alpha(\mathrm{ad}\,u_\beta)^{p^k} = \sum_{i=0}^m \{f(\alpha_i, \beta_i)^{p^k} - f(\alpha_i, \beta_i)^{p^{k-1}} f(\delta_i, \beta_i)^{p^{k-1}(p-1)}\}u_\alpha.$$

Suppose that $\alpha = \Sigma_{i=0}^m \alpha_i = \Sigma_{i=1}^m (\Sigma_{j=1}^{n_i-1} s_{ij}(\alpha)\sigma_{ij} + s_i(\alpha)\delta_i) + \Sigma_{j=1}^{n_0} s_{0j}(\alpha)\sigma_{0j}$. Then

$$\sum_{i=0}^m \{f(\alpha_i, \beta_i)^{p^k} - f(\alpha_i, \beta_i)^{p^{k-1}} f(\delta_i, \beta_i)^{p^{k-1}(p-1)}\}$$

$$= \sum_{i=0}^m \left(\sum_{j=1}^{q_i} s_{ij}(\alpha)\{f(\sigma_{ij}, \beta_i)^{p^k} - f(\sigma_{ij}, \beta_i)^{p^{k-1}} f(\delta_i, \beta_i)^{p^{k-1}(p-1)}\}\right.$$

$$\left. + s_i(\alpha)\{f(\delta_i, \beta_i)^{p^k} - f(\delta_i, \beta_i)^{p^{k-1}} f(\delta_i, \beta_i)^{p^{k-1}(p-1)}\}\right),$$

or

$$\mathrm{ad}\,u_\beta^{p^k} = \sum_{i=0}^m \left(\sum_{j=1}^{q_i} \{f(\sigma_{ij}, \beta_i)^{p^k}\right.$$

$$\left. - f(\sigma_{ij}, \beta_i)^{p^{k-1}} f(\delta_i, \beta_i)^{p^{k-1}(p-1)}\}D(\sigma_{ij}, 0)\right), \tag{1}$$

where $q_0 = n_0$ and $q_i = n_i - 1$ for $i = 1, \ldots, m$. The restricted algebra $\mathfrak{R}(\mathfrak{L}(G, \delta, f))$ is therefore contained within the span of $R \cup T$. In the following discussion we will show that a basis for $\mathfrak{R}(\mathfrak{L}(G, \delta, f))$ is $R \cup T \setminus \{\text{ad } u_{-\delta}\}$ when $G_0 \neq \{0\}$ and $R \cup T \setminus \{\text{ad } u_{-\delta}, D(\delta, 0)\}$ when $G_0 = \{0\}$. It follows that $\dim \mathfrak{R}(\mathfrak{L}(G, \delta, f)) = \dim \mathfrak{L}(G, \delta, f) + n - 2m$.

Definition. The column rank over Φ_p of a matrix A with entries from Φ is the dimension of the vector space over Φ_p spanned by the columns of A. Denote this dimension by col $\text{rank}_{\Phi_p}(A)$.

Lemma 2. *If* $G = G_0$, *then* col $\text{rank}_{\Phi_p}(f(\sigma_{0i}, \sigma_{0j})^p) = n$, *the dimension of* G *over* Φ_p.

Proof. Suppose col $\text{rank}_{\Phi_p}(f(\sigma_{0i}, \sigma_{0j})^p) < n$, that is, suppose that there exist elements $a_1, a_2, \ldots, a_n \in \Phi_p$, not all zero, such that

$$\sum_{j=1}^{n} a_j f(\sigma_{0i}, \sigma_{0j})^p = 0$$

for $i = 1, \ldots, n$. Then from the biadditivity of f we conclude that $f(\sigma_{0i}, \Sigma_{j=1}^n a_j \sigma_{0j})^p = 0$, or $f(\sigma_{0i}, \Sigma_{j=1}^n a_j \sigma_{0j}) = 0$ for $i = 1, \ldots, n$. This contradicts the nondegeneracy of f, whence the lemma is proved.

Lemma 3. *Suppose* $G = G_1$. *Let* $\{\beta_1, \ldots, \beta_k, \delta\}$ *be a basis for* G *where* $f(\beta_1, \delta) \neq 0$, *and let*

$$A = \begin{bmatrix} 0 & f(\beta_1, \beta_2)^p - f(\beta_1, \beta_2) f(\beta_1, \delta)^{p-1} \cdots f(\beta_1, \beta_k)^p - f(\beta_1, \beta_k) f(\beta_1, \delta)^{p-1} \\ \vdots & \vdots \\ f(\beta_k, \beta_1)^p - f(\beta_k, \beta_1) f(\beta_k, \delta)^{p-1} \cdots f(\beta_k, \beta_{k-1}) - f(\beta_k, \beta_{k-1}) f(\beta_k, \delta)^{p-1} & 0 \\ f(\delta, \beta_1)^p \quad \cdot \quad \cdot \quad \cdot & f(\delta, \beta_k)^p \end{bmatrix}$$

Then col $\text{rank}_{\Phi_p}(A) = k$.

Proof. Suppose col $\text{rank}_{\Phi_p}(A) < k$. Then there exist $a_1, \ldots, a_k \in \Phi_p$, not all zero, such that

$$\sum_{j=1}^{k} a_j \{f(\beta_i, \beta_j)^p - f(\beta_i, \beta_j) f(\beta_i, \delta)^{p-1}\} = 0$$

for $i = 1, \ldots, k$ and $\Sigma_{j=1}^k a_j f(\delta, \beta_j)^p = 0$. For each of the first k equalities we have

$$f\left(\beta_i, \sum_{j=1}^{k} a_j\beta_j\right)^p = f\left(\beta_i, \sum_{j=1}^{k} a_j\beta_j\right) f(\beta_i, \delta)^{p-1}$$

or

$$\left(f\left(\beta_i, \sum_{j=1}^{k} a_j\beta_j\right) \Big/ f(\beta_i, \delta)\right)^{p-1} = 1,$$

if $f(\beta_i, \delta) \neq 0$. Thus, for each $i = 1, \ldots, k$, there exists $c_i \in \Phi_p$ such that $f(\beta_i, \Sigma_{j=1}^{k} a_j\beta_j) = c_i f(\beta_i, \delta)$ (if $f(\beta_i, \delta) = 0$, then $c_i = 0$). Now define $g : G \to \Phi$ and $h : G \to \Phi$ by $g(\alpha) = f(\alpha, \delta)$ and $h(\alpha) = f(\beta_1, \alpha)[f(\beta_1, \delta)]^{-1}$, whence $f(\alpha, \beta) = g(\alpha)h(\beta) - g(\beta)h(\alpha)$. Now

$$\begin{aligned} h\left(\sum_{j=1}^{k} a_j\beta_j - c_1\delta\right) &= f\left(\beta_1, \sum_{j=1}^{k} a_j\beta_j\right)[f(\beta_1, \delta)]^{-1} - f(\beta_1, c_1\delta)[f(\beta_1, \delta)]^{-1} \\ &= \left\{ f\left(\beta_1, \sum_{j=1}^{k} a_j\beta_j\right) - c_1 f(\beta_1, \delta)\right\}[f(\beta_1, \delta)]^{-1} = 0 \end{aligned}$$

and

$$g\left(\sum_{j=1}^{k} a_j\beta_j - c_1\delta\right) = f\left(\sum_{j=1}^{k} a_j\beta_j - c_1\delta, \delta\right) = \sum_{j=1}^{k} a_j f(\beta_j, \delta).$$

But $\Sigma\, a_j f(\delta, \beta_j)^p = 0$, so $g(\Sigma_{j=1}^{k} a_j\beta_j - c_1\delta) = 0$. This implies the contradiction $f(\alpha, \Sigma_{j=1}^{k} a_j\beta_j - c_1\delta) = 0$ for all α, implying that col $\text{rank}_{\Phi_p}(A) = k$.

Now suppose G is arbitrary and $\{\sigma_{01}, \ldots, \sigma_{0n_0}, \sigma_{11}, \ldots, \sigma_{1,n_1-1}, \delta_1, \ldots, \sigma_{m1}, \ldots, \delta_m\}$ is a basis of G over Φ_p. Equation (1) shows that $\mathfrak{R}(\mathfrak{L}(G, \delta, f)) \subseteq \langle D(\sigma_{ij}, 0), \text{ad } x \mid x \in \mathfrak{L}(G, \delta, f)\rangle$ and for the special case $k=1$ we have a matrix equation of the form:

$$\begin{bmatrix} \text{ad } u^p_{\sigma_{01}} \\ \vdots \\ \text{ad } u^p_{\sigma_{0n_0}} \\ \vdots \\ \text{ad } u^p_{\sigma_{m1}} \\ \vdots \\ \text{ad } u^p_{\delta_m} \end{bmatrix} = \begin{bmatrix} C_0 & & 0 \\ & \ddots & \\ 0 & & C_m \end{bmatrix} \begin{bmatrix} D(\sigma_{01}, 0) \\ \vdots \\ D(\sigma_{0n_0}, 0) \\ \vdots \\ D(\sigma_{m1}, 0) \\ \vdots \\ D(\sigma_{m,n_m-1}, 0) \end{bmatrix},$$

where $C_0 = (f(\sigma_{0i}, \sigma_{0j})^p)$ and, for $i > 0$, C_i is an $n_i \times (n_i - 1)$ matrix of the form of the matrix in Lemma 3. Denote this matrix by C. To determine the coefficient matrix of the $D(\sigma_{ij}, 0)$'s for higher powers of p, one merely raises the elements in C to the appropriate pth power.

Lemma 4. *Let $A = (a_{ij})$ be an $r \times s$ matrix over a field Φ of characteristic $p > 0$ and let $A_{p^t} = (a_{ij}^{p^t})$ for $t \geq 0$. If* col rank$_{\Phi_p} A = s$, *then* rank$_\Phi (A\, A_p \cdots A_{p^t})^T = s$ *for sufficiently large t (^{T} denotes transpose).*

Proof. Since rank$_\Phi (A \cdots A_{p^i})^T \leq$ rank$_\Phi (A \cdots A_{p^{i+1}})^T \leq s$ for all i there exists some t such that rank$_\Phi (A \cdots A_{p^t})^T =$ rank$_\Phi (A \cdots A_{p^{t+1}})^T$. If rank$_\Phi (A \cdots A_{p^t})^T < s$, then there exist $b_1, \ldots, b_s \in \Phi$, not all zero, such that $\Sigma_{j=1}^s b_j a_{ij}^{p^v} = 0$ for all i, $1 \leq i \leq r$, and all v, $0 \leq v \leq t$. Note that this, and the choice of t, implies $\Sigma_{i=1}^s b_j a_{ij}^{p^{t+1}} = 0$. Assume that the b's have been chosen so that the number of nonzero b_j is minimal. In addition, assume $b_1 = 1$. Then

$$0 = \left(\sum_{j=1}^{s} b_j a_{ij}^{p^v} \right)^p = \sum_{j=1}^{s} b_j^p a_{ij}^{p^{v+1}} .$$

On the other hand, since $\Sigma_{j=1}^s b_j a_{ij}^{p^{v+1}} = 0$ for $0 \leq v \leq t$ we have

$$\sum_{j=1}^{s} (b_j^p - b_j) a_{ij}^{p^{v+1}} = 0.$$

Extracting pth roots and using the minimality of the b's (recall $b_1 = 1$) gives $b_j^p - b_j = 0$ for all j, that is, $b_j \in \Phi_p$ for all j. This contradicts the assumption that col rank$_{\Phi_p} A = s$.

Returning to C, recall that the nondegeneracy of f guarantees that col rank$_{\Phi_p} C_0 = n_0$ and col rank$_{\Phi_p} C_i = n_i - 1$ for $i > 0$. Lemma 4 then allows us to conclude that

$$\langle D(\sigma_{ij}, 0), \text{ad } x \mid x \in \mathfrak{L}(G, \delta, f) \rangle \subseteq \mathfrak{R}(\mathfrak{L}(G, \delta, f)).$$

Inclusion in the other direction was illustrated earlier, completing the proof of the main theorem.

Theorem. *Let $\mathfrak{L}(G, \delta, f)$ be a simple Lie algebra of Block. Then* $\dim \mathfrak{R}(\mathfrak{L}(G, \delta, f)) = \dim \mathfrak{L}(G, \delta, f) + n - 2m$.

Corollary. *Two algebras of Block of the same dimension, $\mathfrak{L}(G, \delta, f)$ and $\mathfrak{L}(G', \delta', f')$, are isomorphic only if $m(G) = m(G')$.*

From the preceding discussion it is evident that for $\mathfrak{L}(G, \delta, f)$ and $\mathfrak{L}(G', \delta', f')$ of the same dimension isomorphism is not guaranteed by the

equality $m(G) = m(G')$. This follows from the fact that $\mathfrak{R}(\mathfrak{L}(G, \delta, f))$ need not be isomorphic to $\mathfrak{R}(\mathfrak{L}(G', \delta', f'))$. For example, let $m(G) = m(G') = 0$ and $n = 4$. Suppose $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ and $\{\beta_1', \beta_2', \beta_3', \beta_4'\}$ are bases for G and G', respectively, where the matrices $(f(\beta_i, \beta_j))$ and $(f(\beta_i', \beta_j'))$ are

$$\begin{bmatrix} 0 & 1 & 0 & x \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -x & 0 & -1 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & x & 0 & 1 \\ -x & 0 & 1 & 0 \\ 0 & -1 & 0 & -1/x \\ -1 & 0 & 1/x & 0 \end{bmatrix},$$

respectively, $x \notin \Phi_p$. In the first case, $\mathfrak{R}(\mathfrak{L}(G, \delta, f)) = \operatorname{ad} \mathfrak{L} + (\operatorname{ad} \mathfrak{L})^p$ while this is not true in the second.

BIBLIOGRAPHY

1. Richard Block, *New simple Lie algebras of prime characteristic*, Trans. Amer. Math. Soc. 89 (1958), 421–449. MR 20 #6446.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF OREGON, EUGENE, OREGON 97403

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

RINGS WITH IDEMPOTENTS IN THEIR NUCLEI

BY

MICHAEL RICH

ABSTRACT. Let R be a prime nonassociative ring. If the set of idempotents of R is a subset of the nucleus of R or of the alternative nucleus of R then it is shown that R is respectively an associative or an alternative ring. Also if R has one idempotent $\neq 0, 1$ which is in the Jordan nucleus or in the noncommutative Jordan nucleus then it is shown that R is respectively a Jordan or a noncommutative Jordan ring.

Introduction. The purpose of this paper is to demonstrate that the degree of associativity of a prime, not necessarily associative ring can be determined from the associativity or lack thereof of the idempotents. Throughout we assume that the ring contains at least one idempotent $\neq 0, 1$. We consider four cases. First, it is easily shown that if R is a prime ring all of whose idempotents lie in the nucleus then R is associative. This motivates consideration of the case in which all of the idempotents lie in an appropriate alternative nucleus of the ring. Similarly, the result here is that the ring is alternative. We next consider a prime commutative ring in which at least one idempotent $\neq 0, 1$ lies in an appropriate Jordan nucleus and show that this implies that the ring is a Jordan ring. Finally, we consider prime flexible rings with at least one idempotent in the appropriate noncommutative Jordan nucleus with the result being that the ring is a noncommutative Jordan ring. Examples are given to show that the conditions assumed are necessary. The latter two cases generalize a result of Osborn.

As usual, the associator (x, y, z) denotes $(xy)z - x(yz)$ and the commutator $[x, y] = xy - yx$. Also R^+ is the same additive group as R, but multiplication in R^+ is given by $a \cdot b = \frac{1}{2}(ab + ba)$, ab being the multiplication in R. Of course, this is meaningful only if $\frac{1}{2}a$ is meaningful for all a in R. A ring is called flexible if $(x, y, x) = 0$, alternative if $(y, x, x) = (x, x, y) = 0$,

---

Received by the editors January 17, 1974.

AMS (MOS) subject classifications (1970). Primary 17D05, 17C99, 17A15, 16A32.

Key words and phrases. Alternative nucleus, Jordan nucleus, noncommutative Jordan nucleus, flexible, idempotent, prime ring.

Jordan if $[x, y] = (x^2, y, x) = 0$ and noncommutative Jordan if it is flexible and $(x^2, y, x) = 0$.

1. The associative case. Let R be an arbitrary nonassociative ring. The nucleus $N(R)$ of R is defined by:

$$N(R) = \{x \in R \mid (x, y, z) = (y, z, x) = (y, x, z) = 0 \quad \forall y, z \in R\}.$$

It is well known [9, p. 13] that $N(R)$ is an associative subring of R.

A ring R is said to have a Peirce decomposition relative to the idempotent $e \in R$ if R can be decomposed into a direct sum of the Z modules R_{ij} $(i, j = 0, 1)$ where $R_{ij} = \{x \in R \mid xe = jx \text{ and } ex = ix\}$. It is known that if R is an associative ring and if e is an idempotent in R then R has a Peirce decomposition relative to e. Also, if R has an identity element 1 and if we write $e_1 = e$ and $e_0 = 1 - e$ then $R_{ij} = e_i R e_j$ [3].

Lemma 1. *Let e be an idempotent of the ring R. Then $e \in N(R)$ if and only if R has a Peirce decomposition $R = \bigoplus R_{ij}$ $(i, j = 0, 1)$ relative to e satisfying the property $R_{ij}R_{kl} \subseteq \delta_{jk}R_{il}$ for $i, j, k, l = 0, 1$ (δ denotes the Kronecker delta).*

Proof. Let $e \in N(R)$. Imbed R into the ring $R' = Z + R$ which contains an identity element 1. Clearly e and $1 - e$ are in $N(R')$. From our earlier remark it follows that $R_{ij} = e_i R e_j$ for $i, j = 0, 1$. Thus $R_{ij}R_{kl} = (e_i R e_j)(e_k R e_l) = e_i R(e_j e_k) R e_l \subseteq \delta_{jk} e_i R e_l = \delta_{jk} R_{il}$.

Conversely, if $R = \bigoplus_{i,j=0} R_{ij}$ such that $R_{ij}R_{kl} \subseteq \delta_{jk}R_{il}$ and $a, b \in R$ then $a = \Sigma_{i,j=0}^1 a_{ij}$, $b = \Sigma_{i,j=0}^1 b_{ij}$. Then $(a, e, b) = \Sigma_{i,j,k,l=0}^1 (a_{ij}, e, b_{kl}) = \Sigma_{i,j,k,l=0}^1 (j - k) a_{ij} b_{kl} = 0$. Similarly $(a, b, e) = (e, a, b) = 0$. Thus, $e \in N(R)$. $\square$

If a ring R contains an idempotent $\neq 0, 1$ and if all the idempotents of R lie in $N(R)$ then we shall call R a nuclear ring.

Theorem 1. *A prime nuclear ring is associative.*

Proof. Let R be a prime nuclear ring with $e \neq 0, 1$ an idempotent of R. By Lemma 1 we have a decomposition $R = \bigoplus R_{ij}$, $i, j = 0, 1$, relative to e with $R_{ij}R_{kl} \subseteq \delta_{jk}R_{il}$. Therefore if $i \neq j$ then $R_{ij}^2 = 0$. Thus, for $i \neq j$, $a_{ij}^2 = 0$ so that $e + a_{ij}$ is an idempotent of R. Since R is nuclear $e + a_{ij} \in N(R)$. But $e \in N(R)$. Therefore $a_{ij} \in N(R)$. Thus $R_{10} + R_{01} \subseteq N(R)$. Since $N(R)$ is a subring of R it follows that $R_{10}R_{01} + R_{01}R_{10} \subseteq N(R)$. This, together with the property $R_{ij}R_{kl} \subseteq \delta_{jk}R_{il}$, allows us to conclude that

$B = R_{10}R_{01} + R_{10} + R_{01} + R_{01}R_{10}$ is an ideal of R contained in $N(R)$. Let $U = \{x \in R | xB = 0\}$. Since $B \subseteq N(R)$ it follows that U is an ideal of R. Since R is a prime ring $UB = 0$ implies $U = 0$ or $B = 0$. But $B = 0$ implies that $R = R_{11} \oplus R_{00}$. Thus, R_{11} and R_{00} are ideals of R such that $R_{11}R_{00} = 0$. From the primeness of R again $R_{11} = 0$ or $R_{00} = 0$. But $e \in R_{11}$ so that $R_{11} \neq 0$. Also $R_{00} = 0$ implies that e is the identity of R contrary to hypothesis. Thus, $B \neq 0$ and $U = 0$. Now, let $r_1, r_2, r_3 \in R$ and $b \in B$. Then, since $b \in N(R)$, $(r_1, r_2, r_3)b = (r_1, r_2, r_3b) \in (R, R, B) \subseteq (R, R, N(R)) = 0$. Therefore, $(R, R, R) \subseteq U = 0$ so that R is an associative ring.

2. The alternative case. Following A. Thedy [10] we define the alternative nucleus $N_A(R)$ of an arbitrary ring R by:

$$N_A(R) = \{r \in R | (x, r, x) = 0 \text{ and } (r, y, x) = (y, x, r) = (x, r, y) \quad \forall x, y \in R\}.$$

If R is 3-torsion free (i.e. if $3a = 0$ for $a \in R$ then $a = 0$) then Thedy has shown that $N_A(R)$ is a subring of R.

Lemma 2. *Let e be an idempotent of a ring R. Then $e \in N_A(R)$ if and only if R has a Peirce decomposition relative to e satisfying the properties:*

(a) $R_{ij}R_{jk} \subseteq R_{ik}$.

(b) $R_{ij}R_{ij} \subseteq R_{ji}$.

(c) $R_{ij}R_{kl} = 0$ *if* $j \neq k$ *and* $(i, j) \neq (k, l)$.

(d) $r_{ij}^2 = 0$ *for any* $r_{ij} \in R_{ij}$, $i \neq j$.

Proof. Let e be an idempotent in $N_A(R)$. Then from the definition of $N_A(R)$ one obtains as in [9, p. 33]

$$e(a_{ij}b_{kl}) = (i + j - k)a_{ij}b_{kl} \tag{1}$$

and

$$(a_{ij}b_{kl})e = (k + l - j)a_{ij}b_{kl}. \tag{2}$$

Thus (a) and (b) follow immediately. Also (d) follows from $(r_{ij}, r_{ij}, e) = 0$ and property (b). To obtain (c) first note that if $x \in R$ such that $xe = sx$ for some $s \in Z$ $(ex = tx$ for some $t \in Z)$ then $s = 0$ or $s = 1$ $(t = 0$ or $t = 1)$. Now in (1) and (2) let $j = 1$ and $k = 0$. Then $e(a_{i1}b_{0l}) = (i + 1)a_{i1}b_{0l}$ and $(a_{i1}b_{0l})e = (l - 1)a_{i1}b_{0l}$. By the preceding remark it follows that $i = 0$ and $l = 1$. Therefore if $a_{i1}b_{0l} \neq 0$ then $(i, j) = (k, l) = (0, 1)$ contrary to hypothesis. The same argument applies if $j = 0$ and $k = 1$.

Conversely, if R has a Peirce decomposition relative to e satisfying (a)–(d) then it is straightforward, using the linearity of the associator, to show that $(x, e, x) = 0$ and $(x, e, y) = (y, x, e) = (e, y, x)$ for arbitrary x, y in R. Thus, $e \in N_A(R)$.

We call a ring R an A-nuclear ring if R contains an idempotent $e \neq 0, 1$ and if every idempotent of R lies in $N_A(R)$.

Henceforth, assume that R is an A-nuclear ring, e an idempotent of R, and $R = R_{11} + R_{10} + R_{01} + R_{00}$ the Peirce decomposition relative to e.

Lemma 3. *The set* $B = R_{10}R_{01} + R_{10} + R_{01} + R_{01}R_{10}$ *is an ideal of* R.

Proof. By Lemma 2 it is sufficient to show that $R_{ii}B + BR_{ii} \subseteq B$ for $i = 0, 1$ which reduces to $R_{ii}(R_{ij}R_{ji}) + (R_{ij}R_{ji})R_{ii} \subseteq B$ for $i \neq j$. Now by (d) of Lemma 2 $a_{ij}^2 = 0$ for $a_{ij} \in R_{ij}$, $i \neq j$. Therefore $e + a_{ij}$ is an idempotent. Hence $a_{ij} \in N_A(R)$ if $i \neq j$ so that $(a_{ii}, a_{ij}, a_{ji}) = -(a_{ii}, a_{ji}, a_{ij})$. Since $i \neq j$ the right-hand side is 0 and we have $(a_{ii}a_{ij})a_{ji} = a_{ii}(a_{ij}a_{ji})$. Thus, $R_{ii}(R_{ij}R_{ji}) = (R_{ii}R_{ij})R_{ji} \subseteq R_{ij}R_{ji} \subseteq B$. Similarly, $(R_{ij}R_{ji})R_{ii} \subseteq B$ so that B is an ideal of R. □

Define $U_i = \{x \in R_{ii} | x(R_{10} + R_{01}) = (R_{10} + R_{01})x = 0\}$ for $i = 0, 1$. Then we have:

Lemma 4. U_i $(i = 0, 1)$ *is an ideal of* R.

Proof. We prove the lemma for U_1 and note that the same proof applies for U_0. Clearly U_1 is an abelian group under addition. Let $u \in U_1$, $a \in R_{10} + R_{01}$, and $r \in R$. Without loss of generality we may assume that $r \in R_{11}$. Also $a \in N_A(R)$. Therefore $(ur)a = u(ra) - (u, a, r)$. Now $ra \in R_{10}$ and $ar \in R_{01}$. Therefore $u(ra) = 0 = u(ar)$. Hence $(ur)a = 0$. Similarly $a(ur) = (au)r + (u, a, r)$ so that $a(ur) = 0$. Therefore $ur \in U_1$. In the same vein $ru \in U_1$. Thus U_1 is an ideal of R.

Lemma 5. $U_iB = BU_i = 0$.

Proof. We again prove the lemma for U_1. Clearly $U_1(R_{10} + R_{01} + R_{01}R_{10}) = (R_{10} + R_{01} + R_{01}R_{10})U_1 = 0$ by Lemma 2 and the definition of U_1. Let $u \in U_1$, $a_{10} \in R_{10}$, and $a_{01} \in R_{01}$. Then since $a_{10} \in N_A(R)$, $u(a_{10}a_{01}) = (ua_{10})a_{01} + (a_{10}, u, a_{01})$. But since $u \in U_1$ the right-hand side is 0. Therefore $u(R_{10}R_{01}) = 0$. Similarly $(R_{10}R_{01})u = 0$. Therefore $U_1B = BU_1 = 0$.

Lemma 6. *If R is a prime A-nuclear ring then R_{11} and R_{00} are associative subrings of R.*

Proof. Since R is a prime ring, by Lemma 5 either $B = 0$ or $U_1 = U_0 = 0$. But $B = 0$ implies that $R = R_{11} \oplus R_{00}$. Since $e \in R_{11}$, $R_{11} \neq 0$. On the other hand, $R_{00} = 0$ implies that $R = R_{11}$ so that e is an identity element of R contrary to hypothesis. Therefore $U_1 = U_0 = 0$. Now let $x, y, z \in R_{ii}$, $r \in R_{10} + R_{01}$. Then $r \in N_A(R)$ and $(x, y, z)r = [(xy)z]r - [x(yz)]r = (xy, z, r) - (x, yz, r) + (xy)(zr) - x[(yz)r] = (xy, z, r) - (x, yz, r) + (x, y, zr) - x(y, z, r) = 0$ since $r, zr \in R_{10} + R_{01} \subseteq N_A(R)$ and if $a, b \in R_{ii}$, $r \in R_{10} + R_{01}$, then $(a, b, r) = (b, r, a) \in [R_{ii}(R_{10} + R_{01})]R_{ii} + R_{ii}[(R_{10} + R_{01})R_{ii}] = 0$. In the same fashion $r(x, y, z) = 0$. Therefore $(x, y, z) \in U_i = 0$. Thus, R_{11} and R_{00} are associative subrings of R.

Theorem 2. *If R is a prime A-nuclear ring then R is alternative.*

Proof. Let $x, y \in R$. Then $x = \Sigma^1_{i,j=0} x_{ij}$ and $y = \Sigma^1_{i,j=0} y_{ij}$ so that $(x, x, y) = \Sigma^1_{i,j=0}(x, x, y_{ij})$. Now if $i \neq j$ then $y_{ij} \in N_A(R)$ so that, by the definition of $N_A(R)$, $(x, x, y_{10}) = (x, x, y_{01}) = 0$. Thus, (x, x, y) reduces to $\Sigma^1_{l=0}(x, x, y_{ll}) = \Sigma^1_{i,j,k,r,l=0}(x_{ij}, x_{kr}, y_{ll})$. Let S denote the sum $\Sigma^1_{i,j,k,r,l=0}(x_{ij}, x_{kr}, y_{ll})$. The terms in S of the form (x_{jj}, x_{kk}, y_{ll}) are all zero by Lemmas 2 and 6. The terms in S of the form (x_{ij}, x_{ij}, y_{ll}) for $i \neq j$ are all zero since $x_{ij} \in N_A(R)$. Finally, the other terms in S come in pairs of the form $(x_{ij}, x_{kr}, y_{ll}) + (x_{kr}, x_{ij}, y_{ll})$. Since $i \neq j$ or $k \neq r$ the sum of each of these pairs is zero. Thus $S = 0$ so that $(x, x, y) = 0$. Similarly $(y, x, x) = 0$. Thus R is alternative. □

It is worthwhile to note that if R is 3-torsion free then Theorem 2 can be obtained more directly. For, in this case, $N_A(R)$ is a subring of R. Thus B is an ideal of R contained in $N_A(R)$. Then by Lemma 3 of [10] $(x, x, y) \in B^\perp$ and $(y, x, x) \in B^\perp$ where $B^\perp = \{r \in R \mid rB = Br = 0\}$. Since $B^\perp$ is an ideal of R and $BB^\perp = B^\perp B = 0$ while $B \neq 0$, it follows that $B^\perp = 0$. Thus R is alternative.

Theorems 1 and 2 assume that R is prime and that all of the idempotents of R lie in $N(R)$, $N_A(R)$, respectively. The following examples show that these conditions are necessary.

Example 1. Let F be a field and R an algebra over F with basis elements e, b, b', c, f with multiplication given by: $e^2 = e$, $f^2 = f$, $eb = bf = b$,

$eb' = b'f = b'$, $ce = fc = c$, $bc = e$, $cb' = f$, and all other products zero. It is straightforward to see that $e \in N(R)$ and that R is a simple algebra, hence a simple ring. However, R is not even alternative since $(b, c, b') + (b', c, b) = b' - b \neq 0$. This is due to the fact that $e + b$ is an idempotent of R but $e + b \notin N_A(R)$. Note also that R does not satisfy the Jordan identity $(x^2, y, x) = 0$ since $((f + b')^2, c, f + b') = -b' \neq 0$.

Example 2. Let R be a 3-dimensional algebra over a field F with basis e, a, b and multiplication given by: $e^2 = e$, $ab = a - b$, $ba = b$, and all other products zero. Then $e \in N(R)$ and e is the only idempotent of R. Thus, R is a nuclear ring. In addition, R is a semiprime ring. However, R is not a prime ring since the ideals Fe and $Fa + Fb$ are orthogonal. R is not alternative since $(a, b, b) = a - b \neq 0$. Thus, the assumption that R is prime is necessary. Here again, R does not satisfy $(x^2, y, x) = 0$.

3. The Jordan case. Henceforth we must assume that all of our rings R satisfy the condition that to each $a \in R$ there exists a unique $b \in R$ such that $2b = a$. We write $b = \frac{1}{2}a$. It is known [2], [4] that if R is a Jordan ring and if e is an idempotent of R then R has a decomposition $R = R_1 + R_{1/2} + R_0$ where $R_i = \{x \in R \mid xe = ex = ix\}$. Also, the modules R_i satisfy the multiplicative properties:

(i) $R_i^2 \subseteq R_i$ for $i = 0, 1$; $R_{1/2}^2 \subseteq R_1 + R_0$, $R_1R_0 = 0$, $R_iR_{1/2} \subseteq R_{1/2}$ for $i = 0, 1$.

Thus, if $a, b \in R_{1/2}$ then $ab \in R_1 + R_0$. We denote this by $ab = (ab)_1 + (ab)_0$. It is also known that products of elements of the different R_i satisfy:

(ii) (a) $x_{1/2}(y_iz_i) = (x_{1/2}y_i)z_i + (x_{1/2}z_i)y_i$, $i = 0, 1$.

(b) $x_i(y_{1/2}z_{1/2}) = [(x_iy_{1/2})z_{1/2} + (x_iz_{1/2})y_{1/2}]_i$, $i = 0, 1$.

(c) $[(x_1y_{1/2})z_{1/2}]_0 = [(x_1z_{1/2})y_{1/2}]_0$.

(d) $[(x_0y_{1/2})z_{1/2}]_1 = [(x_0z_{1/2})y_{1/2}]_1$.

(e) $(x_1y_{1/2})z_0 = x_1(y_{1/2}z_0)$.

We define the Jordan nucleus, $N_J(R)$, of a commutative ring R by:

$$N_J(R) = \{a \in R \mid (ab)(cd) + (ad)(bc) + (ac)(bd) = [b(cd)]a + [b(ac)]d + [b(ad)]c$$
$$= [a(bc)]d + [a(bd)]c + [a(cd)]b \text{ for all } b, c, d \in R\}.$$

Thus, an element $a \in R$ is in $N_J(R)$ if it satisfies the linearized version of the Jordan identity.

Lemma 7. *Let e be an idempotent of a commutative ring R. Then $e \in N_J(R)$ if and only if the elements of the spaces R_i relative to e satisfy* (i) *and* (ii).

Proof. (i) and (ii) are established for Jordan rings in [4]. Since the procedure in all cases is to linearize the Jordan identity and to specialize by setting one of the elements equal to e we may conclude immediately that $e \in N_J(R)$ implies (i) and (ii).

One may verify directly that if (i) and (ii) are satisfied then $e \in N_J(R)$ by setting $a = e$ in the definition of $N_J(R)$ and decomposing b, c and d into their components. The proof is straightforward but the computations are lengthy. We do not present the computations here. □

If R is a commutative ring with at least one idempotent $e \neq 0, 1$ lying in $N_J(R)$ then we call R a J-nuclear ring. Osborn has shown [6], [7, Proposition 6.7] that if R is a commutative ring satisfying (i) and (ii) then R is a Jordan ring if and only if R_1 and R_0 are Jordan rings. Thus if R is simple then R is Jordan. The following theorem draws from and generalizes Osborn's result.

Theorem 3. *If R is a prime J-nuclear ring then R is a Jordan ring.*

Proof. Let e be an idempotent $\neq 0, 1$ in R such that $e \in N_J(R)$. By Lemma 7 we have (i) and (ii). Let $A = (R_{1/2}R_{1/2})_1 + R_{1/2} + (R_{1/2}R_{1/2})_0$. It follows from (i) and (ii)(b) that A is an ideal of R. Also, let $C_i = \{x \in R_i | xR_{1/2} = 0\}$ for $i = 0, 1$. It follows from (i) and (ii)(a) that C_i, $i = 0, 1$, is an ideal of R. Also, from (i) and (ii)(b) $AC_1 = AC_0 = 0$. Since R is a prime ring either $A = 0$ or $C_1 = C_0 = 0$. But $A = 0$ implies that $R = R_1 \oplus R_0$. This, however, is impossible as in Theorem 2. Therefore $C_1 = C_0 = 0$.

From (ii)(a) we have a homomorphism ϕ_i from R_i into $\mathrm{Hom}(R_{1/2}, R_{1/2})^+$ with $\mathrm{Ker}\,\phi_i = C_i$ for $i = 0, 1$ [2]. Since $C_1 = C_0 = 0$ we have R_1 and R_0 imbedded in the Jordan ring $\mathrm{Hom}(R_{1/2}, R_{1/2})^+$. Therefore R_1 and R_0 are Jordan rings and by [7, Proposition 6.7] it follows that R is a Jordan ring.

4. The noncommutative Jordan case. Recall that a ring R is a noncommutative Jordan ring if it is flexible and satisfies the identity $(x^2, y, x) = 0$. It is known [1], [5] that if e is an idempotent of a noncommutative Jordan ring R then R has a decomposition $R = R_1 + R_{1/2} + R_0$ relative to e where $R_i = \{x \in R | xe = ex = ix\}$ for $i = 0, 1$ and $R_{1/2} = \{x \in R | xe + ex = x\}$. Multiplication among the R_i is given by:

(iii) $R_i^2 \subseteq R_i$, $R_iR_{1/2} + R_{1/2}R_i \subseteq R_{1/2}$, $i = 0, 1$, $R_1R_0 = R_0R_1 = 0$ and if $x, y \in R_{1/2}$ then $xy + yx \in R_1 + R_0$.

Assume now that R is a flexible ring in which for every a in R there is a unique b in R such that $2b = a$. We define the noncommutative Jordan

nucleus, $N_{NJ}(R)$, of R by:

$$N_{NJ}(R) = \{a \in R \mid [E_{ax+xa}, F_z] + [E_{az+za}, F_x] + [E_{xz+zx}, F_a] = 0$$
$$= a([E_{xy+yx}, F_z] + [E_{xz+zx}, F_y] + [E_{zy+yz}, F_x])$$
$$\text{for all } x, y, z \text{ in } R\}$$

where $E, F = r, l$ and r_x (l_x) denotes right (left) multiplication by the element x. It is a straightforward matter to show that $N_{NJ}(R) \subseteq N_J(R^+)$. The properties (iii) are obtained for noncommutative Jordan rings by linearizing the Jordan identities and setting one of the variables equal to e. McCrimmon [5] has shown by the same method that

(3) $\qquad e(zy + yz) = zy, \quad (yz + zy)e = yz \quad$ if $y \in R_0$ and $z \in R_{1/2}$

and

(4) $\qquad zl_{xy} = zl_y l_x + zr_x l_y \quad$ and $\quad zr_{xy} = zl_y r_x + zr_x r_y$

if $x, y \in R_0$ and $z \in R_{1/2}$.

If R contains an idempotent $e \neq 0, 1$ such that $e \in N_{NJ}(R)$ then we shall call R an NJ-nuclear ring. Thus, in an NJ-nuclear ring (iii), (3) and (4) hold. Similarly, we have

(3′) $\qquad e(zy + yz) = yz, \quad (yz + zy)e = zy \quad$ if $y \in R_1$ and $z \in R_{1/2}$

and

(4′) $\qquad zl_{xy} = zl_y l_x + zl_x r_y, \quad zr_{xy} = zr_x r_y + zr_y l_x$

if $x, y \in R_1$ and $z \in R_{1/2}$.

For, since R is flexible, $l_{ay} - l_y l_a = r_{ya} - r_y r_a$ for all $a, b \in R$. In particular, if $a = e$, $y \in R_1$ and we allow this to act on $z \in R_{1/2}$ we get $yz - e(yz) = zy - (zy)e$ or $yz - zy = e(yz) - (zy)e$. Add and subtract $e(zy)$ to the right side of this equation to get $yz - zy = e(yz) + e(zy) - zy$. Therefore $yz = e(yz) + e(zy)$. The second half of (3′) follows in a similar manner.

For the first half of (4′) let $E = l$, $F = r$, $a = e$, $x \in R_1$, and $z \in R_{1/2}$ in the definition of $N_{NJ}(R)$ to get $2[l_x, r_z] + [l_z, r_x] + [l_{xz+zx}, r_e] = 0$ which, by flexibility, reduces to $[l_x, r_z] + [l_{xz+zx}, r_e] = 0$. If we allow this to act on $y \in R_1$ we get $(xy)z - x(yz) + [(xz + zx)y]e - (xz + zx)y = 0$. Again, in the definition of $N_{NJ}(R)$ let $x \in R_1$, $a = z = e$ to obtain $2[E_x, F_e] + 2[E_e, F_x] + 2[E_x, F_e] = 0$. By flexibility $[E_x, F_e] + [E_e, F_x] = 0$. There-

fore, we have

$$[E_x, F_e] = 0 \text{ if } x \in R_1. \tag{5}$$

Therefore $[(xz + zx)y]e = [(xz + zx)e]y = (zx)y$ by (3′). Thus, we now have $(xy)z - x(yz) - (xz)y = 0$ which reduces to $zl_{xy} = zl_y l_x + zl_x r_y$. Similarly if we let $E = r$, $F = l$, $x \in R_1$, $z \in R_{1/2}$, and $a = e$ we get the second half of (4′).

Lemma 8. *Let R be an NJ-nuclear ring with* $K_i = \{x \in R_i | xR_{1/2} = R_{1/2}x = 0\}$ *for* $i = 0, 1$. *Then* K_i *is an ideal of R.*

Proof. If $i = 0$ this follows from (iii) and (4) while if $i = 1$ it follows from (iii) and (4′).

Lemma 9. *If R is an NJ-nuclear ring and* $C_i = \{x \in R_i | x \cdot R_{1/2} = 0\}$, $i = 0, 1$ *then* $K_i = C_i$.

Proof. Clearly $K_i \subseteq C_i$. Let $y \in C_i$, $z \in R_{1/2}$. Then $yz + zy = 0$. Then if $i = 0$, (3) gives $yz = zy = 0$; whereas, if $i = 1$, one gets the same result from (3′). Thus, $y \in K_i$. □

We have noted earlier that if $x \in R_1$ and $z \in R_{1/2}$ in an NJ-nuclear ring then $[l_x, r_z] + [l_{xz+zx}, r_e] = 0$. From flexibility we also get $[l_z, r_x] + [l_e, r_{xz+zx}] = 0$. If we allow these to act on $y \in R_{1/2}$ we obtain:

$$(xy)z - x(yz) + [(xz + zx)y]e - (xz + zx)(ye) = 0 \tag{6}$$

and

$$(zy)x - z(yx) + (ey)(xz + zx) - e[y(xz + zx)] = 0, \tag{7}$$

if $y, z \in R_{1/2}$ and $x \in R_1$.

Similarly, if $x \in R_0$, $z \in R_{1/2}$ and $a = e$ in the definition of $N_{NJ}(R)$ we get $[E_z, F_x] + [E_{xz+zx}, F_e] = 0$. If we allow this to act on $y \in R_{1/2}$ we obtain:

$$x(yz) - (xy)z + e[y(xz + zx)] - (ey)(xz + zx) = 0 \tag{6′}$$

and

$$(zy)x - z(yx) + [(xz + zx)y]e - (xz + zx)(ye) = 0, \tag{7′}$$

if $y, z \in R_{1/2}$ and $x \in R_0$.

We are now able to prove:

Theorem 4. *A prime NJ-nuclear ring is a noncommutative Jordan ring.*

Proof. We first show that $K_0 = K_1 = 0$. As in Theorem 3 let $A =$

$(R_{½}R_{½})_0 + R_{½} + (R_{½}R_{½})_1$. Then as in [5, Lemma 2] A is an ideal of R. Now by (iii), (6), and (7), $AK_1 = K_1A = 0$; whereas by (iii), (6′), and (7′), $AK_0 = K_0A = 0$. Now, if $A = 0$ then $R_{½} = 0$ which is impossible since R is a prime ring. Therefore $K_1 = K_0 = 0$. Thus, by Lemma 9, $C_1 = C_0 = 0$. Therefore R^+ is a J-nuclear ring in which $C_1 = C_0 = 0$. As in Theorem 3 it follows that R_1^+ and R_0^+ are Jordan rings. Therefore by [6], [7], R^+ is a Jordan ring. Since R is flexible and R^+ is Jordan, it follows [8] that R is a noncommutative Jordan ring.

Finally, note that our Example 1 earlier shows that it is not true that in a prime nonflexible ring $e \in N_{NJ}(R)$ implies that $R = N_{NJ}(R)$.

REFERENCES

1. A. A. Albert, *Power-associative rings*, Trans. Amer. Math. Soc. 64 (1948), 552–593. MR 10, 349.

2. ———, *A theory of power-associative commutative algebras*, Trans. Amer. Math. Soc. 69 (1950), 503–527. MR 12, 475.

3. N. Jacobson, *Structure of rings*, 2nd rev. ed., Amer. Math. Soc. Colloq. Publ., vol. 37, Amer. Math. Soc., Providence, R. I., 1964. MR 36 #5158.

4. ———, *Structure and representations of Jordan algebras*, Amer. Math. Soc. Colloq. Publ., vol. 39, Amer. Math. Soc., Providence, R. I., 1968. MR 40 #4330.

5. K. McCrimmon, *Structure and representations of noncommutative Jordan algebras*, Trans. Amer. Math. Soc. 121 (1966), 187–199. MR 32 #5700.

6. J. M. Osborn, *Commutative algebras satisfying an identity of degree four*, Proc. Amer. Math. Soc. 16 (1965), 1114–1120. MR 32 #2451.

7. ———, *Varieties of algebras*, Advances in Math. 8 (1972), 163–369. MR 44 #6775.

8. R. D. Schafer, *Noncommutative Jordan algebras of characteristic* 0, Proc. Amer. Math. Soc. 6 (1955), 472–475. MR 17, 10.

9. ———, *An introduction to nonassociative algebras*, Pure and Appl. Math., vol. 22, Academic Press, New York and London, 1966. MR 35 #1643.

10. A. Thedy, *On rings with completely alternative commutators*, Amer. J. Math. 93 (1971), 42–51. MR 43 #300.

DEPARTMENT OF MATHEMATICS, TEMPLE UNIVERSITY, PHILADELPHIA, PENNSYLVANIA 19122

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

ON THE EXTENSION OF MAPPINGS IN STONE-WEIERSTRASS SPACES[1]

BY

ANTHONY J. D'ARISTOTLE

Dedicated to Orrin Frink, Jr.

ABSTRACT. N. Veličko generalized the well-known result of A. D. Taĭmanov on the extension of continuous functions by showing that Taĭmanov's theorem holds when Y (the image space) is H-closed and Urysohn and the mapping f is weakly θ-continuous. We obtain, in a more direct fashion, an even stronger generalization of this theorem.

We proceed to show that the class of all SW spaces is not reflective in the category of all completely Hausdorff spaces and continuous mappings. However, an epi-reflective situation is achieved by suitably enlarging the class of admissible morphisms.

We conclude by establishing a number of results about SW extension spaces.

1. Preliminaries. A subset A of a space X is said to be a zero-set of X if there exists a function f in $C(X)$ (the set of all continuous, real-valued functions on X) such that $A = f^{-1}(\{0\}) = \{x \in X: f(x) = 0\}$. Complements of zero-sets are called cozero-sets. If $C^*(X)$ is the set of all bounded functions in $C(X)$, the subset A of X is said to be C^*-embedded in X if every function in $C^*(A)$ can be extended to a function in $C^*(X)$ [11]. If X and Y are spaces, let $C(X, Y)$ denote the set of all continuous functions from X to Y.

A mapping f of the space X into the space Y is said to be θ-continuous (weakly θ-continuous) if for an arbitrary point $x \in X$ and an arbitrary open set V of Y containing $y = f(x)$, there exists an open set U of X with $x \in U$ and $f(\mathrm{cl}_X U)(f(U)) \subseteq \mathrm{cl}_Y V$ ([8], [19]). The point x is a member of the θ-closure of the set S in X if and only if $(\mathrm{cl}_X G) \cap S \neq \emptyset$ for all open sets G containing x. The set S is said to be θ-closed if it is equal to its θ-closure [25].

A space X is said to be quasicompact if every family of zero-sets of X with the finite intersection property has a nonempty intersection [9]. The space X is said to be Urysohn if distinct points of X are contained in disjoint closed neighborhoods, and is said to be completely Hausdorff if for

---

Received by the editors August 16, 1972 and, in revised form, March 11, 1974.
AMS (MOS) subject classifications (1970). primary 54C20, 54C30, 54C45.

(1) Portions of this paper were presented to the Second Pittsburgh International Conference on General Topology and its Applications, December 18–22, 1972.

every pair x, y of distinct points there is a function f in $C(X)$ such that $f(x) \neq f(y)$. The word *space*, unqualified, shall henceforth mean a completely Hausdorff space. As in [21], the space X is said to be a Stone-Weierstrass space (or briefly, an SW space) if every point-separating subalgebra of $C^*(X)$ which contains the constants is uniformly dense in $C^*(X)$.

A mapping $f: X \to Y$, where X, Y are arbitrary spaces, is said to be cozero-set continuous if $f^{-1}(C)$ is open for all cozero-sets C of Y. Cozero-set continuous functions will be referred to as c-maps. If X is a space, let $\tilde{X}$ be the space obtained by taking on the same set X the weak topology relative to $C(X)$. As noted in [3], X is SW if and only if $\tilde{X}$ is compact. Moreover, a map f from an arbitrary space X to Y is a c-map if and only if $\gamma_Y f: X \to \tilde{Y}$ is continuous, where $\gamma_Y: Y \to \tilde{Y}$ is the identity map.

A Hausdorff space X is said to be absolutely closed, or simply H-closed, if it is closed in every Hausdorff space in which it can be embedded. This concept is a generalization of a property of compact Hausdorff spaces, and was introduced in 1924 by Alexandroff and Urysohn [1]. In [15], Katětov showed that any Hausdorff space X could be densely embedded in an H-closed space κX, now referred to as the Katětov extension of X, having the property that X is a C^*-embedded subset. For a construction of κX, the reader is referred to [18].

A filter $\mathcal{F}$ on a space X is said to be completely regular if $\mathcal{F}$ has a base $\mathcal{B}$ of open sets such that for each set $A \in \mathcal{B}$ there exist a set $B \in \mathcal{B}$ contained in A and a function $f \in C(X)$ which is equal to 0 on B and 1 on $X \setminus A$ [5]. The filter $\mathcal{F}$ is said to be free or fixed according as the intersection of all its members is empty or nonempty.

The space Y is said to be an extension of the space X if there exists a homeomorphism h from X into Y such that $h(X)$ is dense in Y. If h is the identity map, the reference to h is omitted. The extensions Y and Z of X are said to be isomorphic if there is a homeomorphism of Y onto Z which leaves X pointwise fixed.

An arbitrary topological space X is said to be realcompact if every real maximal ideal in $C(X)$ is fixed [6].

An open filter is a filter in the lattice of open sets. The open filter $\mathcal{A}$ is said to have the countable closure intersection property (abbreviated c.c.i.p.) provided that for each countable subset $\mathcal{C}$ of $\mathcal{A}$, $\bigcap\{\mathrm{cl}_X C: C \in \mathcal{C}\} \neq \emptyset$. An open ultrafilter is a maximal open filter.

A Hausdorff space X is said to be almost realcompact if every open ultrafilter with the c.c.i.p. converges.

2. Extension of maps. The following theorem is proved by Taĭmanov in [23].

(2.1) *Let A be a dense subspace of an arbitrary space X, and let f: $A \rightarrow Y$ be a continuous mapping of A into the compact Hausdorff space Y. The mapping f has a continuous extension from X to Y if and only if, for every pair F_1, F_2 of closed disjoint subsets of Y, we have* $\mathrm{cl}_X f^{-1}(F_1) \cap \mathrm{cl}_X f^{-1}(F_2) = \emptyset$.

It is easily verified that in the above theorem we may replace "closed disjoint subsets" by "disjoint zero-sets".

Lemma 2.2. *Let f: $X \rightarrow Y$ be a map from an arbitrary space X to an H-closed Urysohn space Y. Then the following conditions are equivalent*:

(a) *f is θ-continuous*,

(b) *f is weakly θ-continuous, and*

(c) *f is a c-map.*

Proof. That (a) implies (b) is obvious. A weakly θ-continuous function is always a c-map and therefore it remains to prove that (c) implies (a).

Suppose $p \in X$ and that $f(p) \in V$, where V is open in Y. An H-closed Urysohn space is an SW space [18] and hence completely Hausdorff. There is for each $q \in Y \setminus \{f(p)\}$ a function h_q in $C(Y)$ with $h_q(q) = 0$ and $h_q(f(p)) = 1$. It is clear that $C_q = \{y \in Y: h_q(y) < \frac{1}{2}\}$ is open in Y, $D_q = \{y \in Y: h_q(y) \leq \frac{1}{2}\}$ is a zero-set of Y, $f(p) \notin D_q$, and $Y \subseteq \bigcup \{C_q : q \in Y \setminus \{f(p)\}\} \cup V$. The space Y is H-closed, and so there exist elements $q_1, q_2, q_3, \cdots, q_n$ of $Y \setminus \{f(p)\}$ with $Y \subseteq \bigcup_{i=1}^n \mathrm{cl}_Y C_{q_i} \cup \mathrm{cl}_Y V = \bigcup_{i=1}^n D_{q_i} \cup \mathrm{cl}_Y V$ [16]. Since $\bigcup_{i=1}^n f^{-1}(D_{q_i}) = E$ is a zero-set of X and $p \in X \setminus E$, there is an open set U of X with $p \in U \subseteq \mathrm{cl}_X U \subseteq X \setminus E$. Clearly $f(\mathrm{cl}_X U) \subseteq f(X \setminus E) \subseteq \mathrm{cl}_Y V$ and thus f is θ-continuous.

The following theorem generalizes and follows from (2.1).

Theorem 2.3. *Let A be a dense subspace of an arbitrary space X, and let f: $A \rightarrow Y$ be a c-map from A to the SW space Y. The mapping f has a c-extension from X to Y if and only if, for every pair F_1, F_2 of disjoint zero-sets of Y, we have* $\mathrm{cl}_X f^{-1}(F_1) \cap \mathrm{cl}_X f^{-1}(F_2) = \emptyset$.

Proof. That the condition is necessary follows from the inclusion $\mathrm{cl}_X f^{-1}(F_i) \subseteq g^{-1}(F_i)$ where g: $X \rightarrow Y$ is a c-extension of f. To see that

the condition is sufficient, we note that disjoint zero-sets of Y are disjoint zero-sets of $\tilde{Y}$ and hence $\gamma_Y f\colon A \to \tilde{Y}$ extends to a continuous map $l\colon X \to \tilde{Y}$ by (2.1). The function $h\colon X \to Y$, defined by $\gamma_Y h = l$, is a c-map and $h|A = f$.

Corollary 2.4. *Let A be a dense subset of an arbitrary space X, and let $f\colon A \to Y$ be weakly θ-continuous where Y is H-closed and Urysohn. Then the following conditions are equivalent:*

(a) *f has a weakly θ-continuous extension from X to Y,*

(b) *for any pair F_1, F_2 of θ-closed disjoint subsets of Y, we have* $\mathrm{cl}_X f^{-1}(F_1) \cap \mathrm{cl}_X f^{-1}(F_2) = \varnothing$,

(c) *for any pair F_1, F_2 of disjoint zero-sets of Y, we have* $\mathrm{cl}_X f^{-1}(F_1) \cap \mathrm{cl}_X f^{-1}(F_2) = \varnothing$.

Proof. (a) $\Rightarrow$ (b): Suppose g is the weakly θ-continuous extension of f. If $p \in \mathrm{cl}_X f^{-1}(F_i)$, then $g(p)$ is, easily, a member of the θ-closure of F_i.

(b) $\Rightarrow$ (c): We need merely observe that zero-sets are θ-closed.

(c) $\Rightarrow$ (a): The function f is a c-map by Lemma 2.2 and therefore has a c-extension g from X to Y by Theorem 2.3. Again, by Lemma 2.2, g is weakly θ-continuous.

In the above corollary, there are simple examples showing that we may not replace "weakly θ-continuous" by "continuous". On the other hand, by Lemma 2.2, "weakly θ-continuous" may be replaced by "θ-continuous" or "c-map".

Veličko generalized Taĭmanov's theorem by showing that, under the hypothesis of Corollary 2.4, (a) is equivalent to (b) [25]. Stephenson [21] has shown that there exists a noncompact regular SW space Y. Since a regular absolutely closed space is compact, Y cannot be absolutely closed. Thus, Y is seen to be an example of an SW space which is not H-closed and so Theorem 2.3 covers a wider class of spaces than Veličko's result.

Almost realcompact spaces were defined and studied by Frolík in [10], where he showed that in many instances they behave much like realcompact spaces. In [7], Engelking gave the analogue of Taĭmanov's theorem for completely regular realcompact spaces: Let A be a dense subspace of an arbitrary topological space X, and let $f\colon A \to Y$ be a continuous function of A into the completely regular realcompact space Y. The mapping f has a continuous extension from X to Y if and only if, for any sequence $\{F_i\}_{i=1}^{\infty}$ of closed subsets of Y such that $\bigcap_{i=1}^{\infty} F_i = \varnothing$, we have $\bigcap_{i=1}^{\infty} \mathrm{cl}_X f^{-1}(F_i) = \varnothing$.

It is a simple matter to establish the counterpart of Theorem 2 for completely Hausdorff realcompact spaces. In the same vein, we have

Theorem 2.5. *Suppose Y is almost realcompact, κY is an SW space, A is dense in the arbitrary space X, and $f\colon A \to Y$ is weakly θ-continuous. The mapping f has a weakly θ-continuous extension from X to Y if and only if for any sequence $\{F_i\}_{i=1}^{\infty}$ of θ-closed subsets of Y such that $\bigcap_{i=1}^{\infty} F_i = \varnothing$, we have $\bigcap_{i=1}^{\infty} \mathrm{cl}_X f^{-1}(F_i) = \varnothing$.*

Proof. To prove that the condition is necessary, we need merely argue as in (a) $\Rightarrow$ (b) of Corollary 2.4. We now establish the sufficiency of the condition and first note that f may be regarded as a weakly θ-continuous function from A to κY. If H_1 and H_2 are disjoint θ-closed subsets of κY, then $H_1 \cap Y$ and $H_2 \cap Y$ are disjoint θ-closed subsets of Y, and therefore $\mathrm{cl}_X f^{-1}(H_1) \cap \mathrm{cl}_X f^{-1}(H_2) = \mathrm{cl}_X f^{-1}(H_1 \cap Y) \cap \mathrm{cl}_X f^{-1}(H_2 \cap Y) = \varnothing$. By virtue of Corollary 2.4, f has a weakly θ-continuous extension g from X to κY. It remains to show that $g(X) \subseteq Y$.

Suppose, on the contrary, that there is a point p of $X \setminus A$ such that $g(p) = \mathcal{A} \in \kappa Y \setminus Y$. Now $\mathcal{A}$ is an open ultrafilter on Y which does not have the c.c.i.p., and so there exists a countable subset $\mathcal{C} = \{G_i\}_{i=1}^{\infty}$ of $\mathcal{A}$ with $\bigcap_{i=1}^{\infty} \mathrm{cl}_Y G_i = \varnothing$. Since κY is H-closed and Urysohn, $\mathrm{cl}_Y G_i$ is θ-closed for all positive integers i [24], and letting $F_i = \mathrm{cl}_Y G_i$ it follows that $\bigcap_{i=1}^{\infty} \mathrm{cl}_X f^{-1}(F_i) = \varnothing$. There exists a positive integer j such that $p \in X \setminus \mathrm{cl}_X f^{-1}(F_j)$, and we observe that $G_j \cup \{\mathcal{A}\}$ is open in κY and contains $\mathcal{A}$. If U is an arbitrary open subset of X which contains p, choose a point s of $[X \setminus \mathrm{cl}_X f^{-1}(F_j)] \cap U \cap A$. Now $s \notin \mathrm{cl}_X f^{-1}(F_j)$ and hence $s \notin f^{-1}(\mathrm{cl}_Y G_j)$ so that $g(s) = f(s) \notin \mathrm{cl}_Y G_j$. It follows that $g(s) \in Y \setminus \mathrm{cl}_Y G_j$, an open subset of κY which does not intersect $G_j \cup \{\mathcal{A}\}$. Hence $g(U) \not\subseteq \mathrm{cl}_{\kappa Y}[G_j \cup \{\mathcal{A}\}]$, and so g is not weakly θ-continuous at p.

Remarks. Porter and Thomas have given necessary and sufficient conditions on X for κX to be an SW space [18].

It is easily verified that Lemma 2.2 is still valid if it is required merely that κY be Urysohn; thus in the above theorem, "weakly θ-continuous" may be replaced by "θ-continuous" or "c-map".

3. Reflectiveness of SW spaces. Many extensions such as the Stone-Čech compactification, the Hewitt realcompactification, and the Banaschewski zero-dimensional compactification have, on account of their similar

mapping properties, been studied from a categorical standpoint and classified as epi-reflections in appropriate categories. For a thorough discussion of this theory, the reader is referred to [13].

We will not distinguish among isomorphic objects of any category, and for any category, 1_X will denote the identity morphism for the object X. For categorical notions not specifically defined, the reader should consult Mitchell [17].

Definition. If $\mathfrak{A}$ is a full subcategory of a category $\mathfrak{B}$ and if for each object X in $\mathfrak{B}$ there exist an object $X_{\mathfrak{A}}$ in $\mathfrak{A}$ and a morphism (resp. epimorphism) $r\colon X \to X_{\mathfrak{A}}$ such that for each object Y in $\mathfrak{A}$ and each morphism $f\colon X \to Y$, there exists a unique morphism $\overline{f}\colon X_{\mathfrak{A}} \to Y$ such that the diagram

$$\begin{tikzcd} X \arrow[rr, "r"] \arrow[dr, "f"'] & & X_{\mathfrak{A}} \arrow[dl, dashed, "\overline{f}"] \\ & Y & \end{tikzcd}$$

is commutative, then $\mathfrak{A}$ is said to be a reflective (resp. an epi-reflective) subcategory of $\mathfrak{B}$ and r is called a reflection morphism (resp. epimorphism) from X to $X_{\mathfrak{A}}$.

In establishing our next theorem, the techniques employed in Theorem 1 of [12] were most useful. We will furthermore lean heavily upon the following modification of Niemytski's classic example [11, 3K]. Let $X = I^2 = \{(x, y)\colon 0 \le x \le 1,\ 0 \le y \le 1\}$ be the unit square with the usual topology τ_1 and let $A = \{(x, 0)\colon (x, 0) \in X\}$. To each $(x, 0) \in A$ define $V_x = \{(x, 0)\} \cup \{(u, v) \in X\colon v > 0 \text{ and } (u - x)^2 + v^2 < (1/4)^2\}$. Let τ_2 be the topology on X generated by the collection of sets $\tau_1 \cup (V_x)_{x \in I}$ as a subbase. It is easily verified that (X, τ_2) is H-closed and Urysohn and hence SW, and that A with the induced topology is discrete.

Theorem 3.1. *Let $\mathfrak{B}$ be the category of all completely Hausdorff spaces and continuous functions, and let $\mathfrak{A}$ be its full subcategory of all SW spaces. Then $\mathfrak{A}$ is not reflective in $\mathfrak{B}$.*

Proof. Suppose, on the contrary, that $\mathfrak{A}$ is reflective in $\mathfrak{B}$. Consider the space (X, τ_2) described above, and let r be the reflection morphism from A to the SW space $A_{\mathfrak{A}}$. If $i\colon A \to X$ is the identity map, there is a unique morphism $\overline{i}\colon A_{\mathfrak{A}} \to X$ such that $i = \overline{i} \circ r$. Since i is a homeomorphism into, it follows readily that r is a homeomorphism into.

Let D_2 be the discrete space composed of two elements, 1 and 2, let $P = X \times D_2$, and for $n = 1, 2$, let $j_n: X \to P$ be the map defined by $j_n(y) = (y, n)$. For each $y \in i(A)$ identify $j_1(y)$ and $j_2(y)$, let Q be the corresponding quotient space, and let χ be the quotient map from P to Q. Now X and D_2 are H-closed and it follows that P and hence Q is H-closed. One can easily check that Q is Urysohn and therefore SW.

It is now possible to proceed exactly as Herrlich and Strecker have done in [12] to show that A is homeomorphic to the SW space $A_{\mathfrak{U}}$: this is a contradiction since A is not quasicompact.

Let α_X be the set of all nonisomorphic SW extensions Y of X such that X is C^*-embedded in Y, and let e_X denote the set of all those members of α_X with the property that each trace filter is completely regular.

Stephenson [21] introduced and studied a particular member of e_X, which we denote by σX. Specifically, if $\mathfrak{M}$ is the set of all free maximal completely regular filters on X, then σX is the space whose points are the elements of $X \cup \mathfrak{M}$ and whose topology is generated by all sets V^* of the form $V \cup \{\mathfrak{F} \in \mathfrak{M} | V \in \mathfrak{F}\}$ for V open in X. The space σX enjoys many of the properties of the Stone-Čech compactification of a Tychonoff space and is, in fact, homeomorphic to βX in case X is completely regular.

In [20], Raha has described an extension of a space X, which we denote by δX, whose points are again the elements of $X \cup \mathfrak{M}$ and whose topology is similar, in construction, to the topology of the Katětov extension. In particular, any set, open in X, is also open in δX, and if $\mathfrak{F} \in \mathfrak{M}$, basic neighborhoods of $\mathfrak{F}$ are sets of the form $G \cup \{\mathfrak{F}\}$ for $G \in \mathfrak{F}$. It is easily verified that $\delta X \in e_X$.

Lemma 3.2. *If $Y \in \alpha_X$ and $f: X \to Z$ is a c-map from X to the SW space Z, then there exists a unique c-map $g: Y \to Z$ with $g|X = f$.*

Proof. Let F_1 and F_2 be disjoint zero-sets of Z. The sets $A_1 = f^{-1}(F_1)$ and $A_2 = f^{-1}(F_2)$ are disjoint zero-sets of X, and so there is an element l of $C^*(X)$ which is 0 on A_1 and 1 on A_2 [11]. Now l can be extended to a function in $C^*(Y)$ which implies that $\mathrm{cl}_Y A_1 \cap \mathrm{cl}_Y A_2 = \emptyset$. By Theorem 2.3, there is a c-map $g: Y \to Z$ with $g|X = f$. The uniqueness of g follows from the uniqueness of $\gamma_Z g$.

Theorem 3.3. *Let $\mathfrak{B}^*$ be the category of all spaces and c-maps. If $\mathfrak{U}^* \subseteq \mathfrak{B}^*$ is the full subcategory of all SW spaces, then the natural mappings $r: X \to \sigma X$ and $r_1: X \to \delta X$ are reflection epimorphisms for $\mathfrak{U}^*$.*

Proof. The composition of c-maps is a c-map, the identity function is a c-map, and therefore $\mathfrak{B}^*$ is a category. Since X is C^*-embedded in σX and δX ([21], [20]), the proof now follows directly from Lemma 3.2.

4. Projective extrema. If Y is an extension space of X, the trace filters of Y are the filters $\mathfrak{N}(y)$, $y \in Y\setminus X$, where $\mathfrak{N}(y)$ is the filter on X generated by the traces $U \cap X$ of the open sets U of Y which contain y. If X is Tychonoff, then the trace filters $\mathfrak{N}(y)$ of βX are precisely the free maximal completely regular filters on X ([2], [5]).

The extension Y of X is said to be projectively larger than the extension Z of X, denoted $Y \geq Z$, if there exists a continuous surjection $f\colon Y \to Z$ which leaves X pointwise fixed. If η is a class of extensions of X, an element Y of η is said to be a projective maximum (resp. projective minimum) if $Y \geq Z$ (resp. $Z \geq Y$) for all Z in η. Projective maximums (resp. projective minimums), if they exist, are unique [4].

We are now in a position to give a simple proof of the following theorem due to Stephenson [21, Theorem 4(vii)].

Theorem 4.1 (Stephenson). *The projective minimum of e_X is σX.*

Proof. If $Y \in e_X$, let $\tilde{\eta}(y)$ denote the trace filter of $\tilde{Y}$ corresponding to $y \in \widetilde{Y\setminus X}$. Noting that $\beta\tilde{X} = \tilde{Y}$ and that $\tilde{\eta}(y)$ is a subset of the completely regular filter $\eta(y)$, we must have $\eta(y) = \tilde{\eta}(y)$. Let $g\colon Y \to \sigma X$ be the function defined by $g(x) = x$ for $x \in X$ and $g(y) = \eta(y) \in \mathfrak{M}$ for $y \in Y\setminus X$. If $V \cup \{\mathfrak{G} \in \mathfrak{M}\colon V \in \mathfrak{G}\} = T$ is a basic open set of σX, then $g^{-1}(T) = V \cup \{y \in Y\setminus X\colon V \in \eta(y)\}$ which is open in Y by Lemma 4.1 of [18]. Thus, g is continuous.

It is clear that $g(Y)$ is quasicompact and thus SW [3]. It follows that $g(Y)$ is closed in σX [21], and therefore g is onto.

Stephenson also proved that the function g in the above theorem is 1-1. From the manner in which we have defined g, this follows immediately from the fact that Y is completely Hausdorff.

Porter and Thomas [18] and Liu [16] have shown that the Katětov extension is a projective maximum in the class of H-closed extensions of a Hausdorff space X. In view of its affinity with the Katětov extension, it is natural to inquire about the role of δX as a projective maximum.

Theorem 4.2. *If $Y \in \alpha_X$, then $\delta X \geq Y$ if and only if $Y \in e_X$.*

Proof. If $\delta X \geq Y$, then for any $y \in Y\setminus X$ there is an $\mathfrak{A} \in \mathfrak{M}$ such that

$\eta(y) \subseteq \mathfrak{A}$. Since $\tilde{\eta}(y)$ is a maximal completely regular filter and $\tilde{\eta}(y) \subseteq \eta(y)$, it follows that $\eta(y) = \mathfrak{A}$. Hence $Y \in e_X$.

On the other hand, suppose $Y \in e_X$. The identity map $i: X \to Y$ is a c-map, δX is a reflection epimorphism for $\mathfrak{U}^*$, and therefore i can be extended to a c-map $f: \delta X \to Y$. We claim that $f(\mathfrak{M}) \subseteq Y \setminus X$. For if $f(\mathfrak{A}) = x \in X$, let $\mathfrak{D}$ be the class of all cozero-sets of X which contain x. Since f is a c-map and X is C^*-embedded in Y, it follows that $\mathfrak{D} \subseteq \mathfrak{A}$. Now $\mathfrak{D}$ is a base for a fixed maximal completely regular filter, and so $\mathfrak{A}$ is fixed which is a contradiction.

It is clear that f is continuous at each point of x. If $f(\mathfrak{A}) = y \in Y \setminus X$, let U be an open set Y which contains y. Since $\eta(y)$ is completely regular, $\tilde{\eta}(y) = \eta(y)$ which entails the existence of a cozero-set C of Y containing y with $C \cap X \subseteq U$. Since f is a c-map, there is a member G of $\mathfrak{A}$ with $f(G \cup \{\mathfrak{A}\}) \subseteq C$. Clearly, $f(G \cup \{\mathfrak{A}\}) \subseteq U$ and so f is continuous at y. That f is onto follows as in Theorem 4.1. Thus $\delta X \geq Y$.

The following corollary is analogous to a result of Banaschewski in [4].

Corollary 4.3. *If X is a space, then $\sigma X \leq Y \leq \delta X$ for all Y in e_X.*

Remarks. The set e_X (and hence also α_X) may have substantial cardinality. We shall make use of the fact that any Hausdorff extension $Y \supseteq X$ such that $\sigma X \leq Y \leq \delta X$ belongs to e_X. Let N be the space of positive integers with the discrete topology, let $Y = N \cup \mathfrak{M}$, and if $\mathfrak{A} \in \mathfrak{M}$, let $\tau_{\mathfrak{A}}$ be the topology on Y generated by the topology of σN together with $\{N \cup \{\mathfrak{A}\}\}$. Clearly $\sigma N \leq (Y, \tau_{\mathfrak{A}}) \leq \delta N$ and so $(Y, \tau_{\mathfrak{A}}) \in e_N$. If $\mathfrak{A}$ and $\mathfrak{B}$ are distinct members of $\mathfrak{M}$, a routine argument shows that $(Y, \tau_{\mathfrak{A}})$ and $(Y, \tau_{\mathfrak{B}})$ are non-isomorphic extensions of N. Finally, since σN is the Stone-Čech compactification of N, it follows that $2^c = \text{card}\ \mathfrak{M} \leq \text{card}\ e_N$ [11].

It is natural to ask if σX is a projective minimum in α_X. We shall answer this question negatively, but we will first need to describe another SW extension of a given completely Hausdorff space X. Let $\mathfrak{W}$ be the set of all free zero-set ultrafilters (see [11]) on X, and let $\pi X = X \cup \mathfrak{W}$. We define a topology for πX by taking as a base for the open sets the family of all sets of the form $G \cup \{\mathfrak{A} \in \mathfrak{W}: \exists A \in \mathfrak{A} \text{ with } A \subseteq G\}$ where G is any open set of X. It is readily verified that πX is SW and that X is a dense, C^*-embedded subset.

Let $I = [0, 1]$, let τ be the usual topology on I, let J be the subset of I consisting of all irrational numbers, and choose disjoint dense subsets J_1,

J_2 of (I, τ) such that $J = J_1 \cup J_2$. Let $I_1 = I \setminus J_2$, let τ_1 be the topology on I_1 induced by τ, let $\mathcal{S}_1$ be the topology on I_1 generated by $\tau_1 \cup \{J_1\}$, and denote the space $(I_1, \mathcal{S}_1)$ by P. A routine argument shows that (I_1, τ_1) and P have the same continuous functions, and so τ_1 is the collection of cozero-sets of both (I_1, τ_1) and P.

Suppose $g \in C(\pi P, \sigma P)$ and $g|X$ is the identity map. Since no cozero-set of P is contained in J_1, the set J_1 is open in σP. However, if $p \in U \subseteq J_1$ where $U \in \mathcal{S}_1$, then U contains a member of a free zero-set ultrafilter on P, and hence $g^{-1}(J_1) = J_1$ is not open in πP. This is a contradiction, and it is now evident that σP is not a projective minimum in α_P.

Clearly $\pi P \in \alpha_P \setminus e_P$, and it follows from Theorem 4.2 that $\delta P \not\geq P$. Therefore δX is not necessarily the projective maximum of α_X, and this disproves a theorem of Raha [20].

Although σX is not in general a projective minimum in α_X, we do have the following result.

Theorem 4.4. *If X is Tychonoff, then σX is the projective minimum in α_X.*

Proof. If $Y \in \alpha_X$, then $\tilde{Y} = \beta X = \sigma X$; hence $\gamma_Y: Y \to \tilde{Y}$ is the desired map.

However, even if X is Tychonoff, δX need not be the projective maximum of α_X. For in [21], Stephenson has described a Tychonoff space X with a one-point noncompact SW extension Y. Moreover, X is C^*-embedded in Y. If $Y = X \cup \{y\}$ and $\eta(y)$ were completely regular, then one could check the various cases to show that Y would be Tychonoff and hence compact ([3], [14]). Hence $Y \in \alpha_X \setminus e_X$.

5. **Acknowledgement.** The author gratefully acknowledges the substantial help given to him by the referee in the form of five pages of constructive criticism. These comments led to new proofs of Theorems 2.3, 3.3, and 4.4, replacing proofs of my own which were either excessively long, unnatural, or both. In addition, the referee made many other valuable suggestions.

REFERENCES

1. P. Alexandroff and P. Urysohn, *Zur theorie der topologischen Räume*, Math. Ann. 92 (1924), 258–266.

2. P. Alexandroff, *Bikompakte Erweiterungen topologischer Räume*, Mat. Sb. (N. S.) 5 (47) (1939), 403–423. (Russian) MR 1, 318.

3. B. Banaschewski, *On the Weierstrass-Stone approximation theorem*, Fund. Math. 44 (1957), 249–252. MR 19, 1182.

4. ———, *Extensions of topological spaces*, Canad. Math. Bull. 7 (1964), 1–22. MR 28 #4501.

5. N. Bourbaki, *Eléments de mathématique*. Part 1. *Les structures fondamentales de l'analyse*. Livre III: *Topologie générale*. Chaps. 4, 5, 6, 7, 8, 9, 10, Actu-

alités Sci. Indust., nos. 1029, 1045, 1084, Hermann, Paris, 1947, 1958, 1961; English transl., Hermann, Paris; Addison-Wesley, Reading, Mass., 1966. MR 9, 261; 30 #3439; 26 #6918; 34 #5044b.

6. R. F. Dickman, Jr., *Compactifications and realcompactifications of arbitrary topological spaces*, Virginia Polytechnic Institute and State University (unpublished).

7. R. Engelking, *Remarks on realcompact spaces*, Fund. Math. 55 (1964), 303–308. MR 31 #4000.

8. S. Fomin, *Extensions of topological spaces*, Ann. of Math. (2) 44 (1943), 471–480. MR 5, 45.

9. Z. Frolík, *Generalisations of compact and Lindelöf spaces*, Czechoslovak Math. J. 9 (84) (1959), 172–217. (Russian) MR 21 #3821.

10. ———, *A generalization of realcompact spaces*, Czechoslovak Math. J. 13 (88) (1963), 127–138. MR 27 #5224.

11. L. Gillman and M. Jerison, *Rings of continuous functions*, University Ser. in Higher Math., Van Nostrand, Princeton, N. J., 1960. MR 22 #6994.

12. H. Herrlich and G. E. Strecker, *H-closed spaces and reflective subcategories*, Math. Ann. 177 (1968), 302–309. MR 38 #2744.

13. H. Herrlich, *On the concept of reflections in general topology*, Contributions to Extension Theory of Topological Structures (Proc. Sympos., Berlin, 1967), Deutsch. Verlag Wissensch., Berlin, 1969, pp.105–114. MR 44 #2210.

14. E. Hewitt, *Certain generalizations of the Weierstrass approximation theorem*, Duke Math. J. 14 (1947), 410–427. MR 9, 95.

15. M. Katětov, *Über H-abgeschlossene und bikompakte Räume*, Časopis Pěst. Mat. Fys. 69 (1940), 36–49. (German) MR 1, 317.

16. C.-T. Liu, *Absolutely closed spaces*, Trans. Amer. Math. Soc. 130 (1968), 86–104. MR 36 #2107.

17. B. Mitchell, *Theory of categories*, Pure and Appl. Math., vol. 17, Academic Press, New York, 1965. MR 34 #2647.

18. J. Porter and J. Thomas, *On H-closed and minimal Hausdorff spaces*, Trans. Amer. Math. Soc. 138 (1969), 159–170. MR 38 #6544.

19. J. Porter and C. Votaw, *H-closed extensions*. II, University of Kansas, 1973 (preprint).

20. A. B. Raha, *On completely Hausdorff-completion of a completely Hausdorff space*, Pacific J. Math. 38 (1971), 161–166. MR 46 #6299.

21. R. M. Stephenson, Jr., *Spaces for which the Stone-Weierstrass theorem holds*, Trans. Amer. Math. Soc. 133 (1968), 537–546. MR 37 #3337.

22. ———, *Product spaces for which the Stone-Weierstrass theorem holds*, Proc. Amer. Math. Soc. 21 (1969), 284–288. MR 40 #3499.

23. A. D. Taĭmanov, *On extension of continuous mappings of topological spaces*, Mat. Sb. (N. S.) 31 (73) (1952), 459–463. (Russian) MR 14, 395.

24. N. V. Veličko, *H-closed topological spaces*, Mat. Sb. 70 (112) (1966), 98–112; English transl., Amer. Math. Soc. Transl. (2) 78 (1968), 103–118. MR 33 #6576.

25. ———, *On extension of mappings of topological spaces*, Sibirsk. Mat. Ž. 6 (1965), 64–69; English transl., Amer. Math. Soc. Transl. (2) 92 (1970), 41–47. MR 31 #4010; 42 #3.

DEPARTMENT OF MATHEMATICS, BROOKLYN COLLEGE (CUNY), BROOKLYN, NEW YORK 11210

Current address: Departamento de Matematicas, Universidad Simon Bolivar, Sartanejas-Baruta 5354, Caracas, Venezuela

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

NEARNESS STRUCTURES AND PROXIMITY EXTENSIONS

BY

M. S. GAGRAT AND W. J. THRON

ABSTRACT. Proximity, contiguity and nearness structures are here studied from a unified point of view. In the discussion the role that grills can play in the theory is emphasized. Nearness structures were recently introduced by Herrlich and Naimpally. Thron pointed out the importance of grills in proximity theory. Nearness structures ν are then used to generate proximity extensions $(\phi, (X^\nu, \Pi^\nu))$ of a given LO-proximity space (X, Π), where $\Pi_\nu = \Pi$. Finally, the relation of the extensions $(\phi, (X^\nu, \Pi^\nu))$ to arbitrary extensions $(i, (Y, \Pi^*))$ is investigated.

1. Introduction. It was shown by Smirnov [16] that EF-proximities can be used to generate all T_2-compactifications of a given Tychonov space. Somewhat later Ivanova and Ivanov [8] introduced contiguity structures and showed that they can be employed to obtain a large class of T_1-compactifications of a given T_1-space. The concept of contiguity was further investigated and slightly modified by Terwilliger [17]. Very recently Herrlich, Naimpally, and Bentley [6], [12], [2] have introduced nearness structures and have applied them, among others, to the study of extensions of spaces. Also recently Thron [19] brought out the importance of grills in proximity theory.

All of these ideas and concepts are brought to bear here on the study of proximity extensions of proximity spaces. We introduce a construction which may have been first suggested by Bentley (see [14]) and is similar to one employed by Herrlich to obtain completions of N-spaces. The construction associates with every nearness structure ν, compatible with the given proximity Π on X, proximity extensions (X^ν, Π^ν) of the original space (X, Π). This is done in §3. That the simple construction of proximity

Received by the editors October 1, 1973 and, in revised form, March 20, 1974.
AMS (MOS) subject classifications (1970). Primary 54E05, 54C20, 54D35.
Key words and phrases. Proximity structures, contiguity structures, nearness structures, proximity mappings, proximity extensions of proximity spaces, strict extensions, principal extensions, trace system, dual trace system, grill, λ-clan, λ-bunch, λ-cluster.

extensions, employed by Leader [10] for EF-proximities cannot be extended to LO-proximities was recently shown by Naimpally and Whitfield [14]. In §4 we investigate to what extent all proximity extensions can be obtained as (X^ν, Π^ν).

In §2 we take another look at the definitions of proximity, contiguity, and nearness. This is done partly to emphasize the importance of the concept of grill, which appears naturally in $\lambda(\mathfrak{A})$ as well as in maximal λ-compatible families (for λ a contiguity or a nearness). We are also able to bring out the similarities as well as the differences between the three types of structures.

A structure λ shall be called *clan (bunch) generated* if it satisfies the condition

$$\mathfrak{A} \in \lambda \Rightarrow \exists \text{ a } \lambda\text{-clan (bunch) } \mathfrak{G} \text{ such that } \mathfrak{A} \subset \mathfrak{G}.$$

It is known [19] that all basic proximities are clan generated. In §2 we show that the same is true for all basic contiguities. Very recently Naimpally and Whitfield [13] have given an example of a nearness which is not clan generated. It follows that in this important respect nearness structures are much more complicated than proximities or contiguities.

Bunch generated structures are exactly the ones which can be topologically induced. A structure λ on X is said to be *topologically induced* if (X, c_λ) can be embedded via a map ϕ in a topological space (Y, d) and $\mathfrak{A} = [A_i : i \in I] \in \lambda$ if $|\mathfrak{A}|$ is appropriately restricted and $\bigcap [d(\phi(A_i)) : i \in I] \neq \emptyset$. Here c_λ is the closure operator induced by λ (see Definition 2.5). For details on this result see Bentley [3].

In what follows there is always an underlying nonempty set X and frequently also a set $Y \supset X$. It will be convenient to denote elements of X or Y by $x, y, \ldots$, subsets by $A, B, \ldots$. Families of subsets will be denoted by $\mathfrak{A}, \mathfrak{B}, \ldots$. In particular $\mathfrak{F}$ will be used for filters, $\mathfrak{U}, \mathfrak{V}$ for ultrafilters, and $\mathfrak{G}$ for grills. Letters $\alpha, \beta, \gamma, \ldots$ shall be used for collections of families of sets (i.e. $\alpha \subset \mathfrak{P}(\mathfrak{P}(X))$). For nearness structures we shall use $\nu, \mu, \ldots$, for contiguities $\xi, \zeta, \ldots$, a collection which may be any of the three structures shall be denoted by $\lambda, \ldots$. However, for proximities we shall continue to use Π.

In analogy to its use for relations we shall employ the notation $\lambda(\mathfrak{A})$ to mean $\lambda(\mathfrak{A}) = [A : [A] \cup \mathfrak{A} \in \lambda]$. In addition $\lambda([[x]])$ shall be simplified to $\lambda(x)$ and $\Pi([A])$ to $\Pi(A)$. Otherwise we shall refrain from using abbreviations. In particular we shall write $\mathfrak{A} \in \lambda$ or $\mathfrak{A} \notin \lambda$. The notation $|A|, |\mathfrak{A}|, \ldots$ refers to the cardinal number of the set under consideration.

Clusters were initially defined for proximities by Leader [10] as λ-closed λ-clans (see Definition 2.8). We use this definition also for contiguities and for nearness structures, for both of which one can prove (Theorems 2.4 and 2.5) that the λ-closed λ-clans are exactly the maximal λ-compatible families. Thus there is no real conflict between our definition and that of Herrlich, Naimpally, and Bentley [6], [12], [2] who, for near structures, define a cluster as a maximal λ-compatible family. In order to save space many straightforward proofs shall be omitted or given only in outline. The authors would also like to thank S. A. Naimpally and the referee for a large number of valuable comments.

2. Proximities, contiguities, and nearness structures. We begin by recalling the definition of a stack and a grill.

Definition 2.1. A family $\mathfrak{S}$ of subsets of X is called a *stack* on X if it satisfies $A \supset B \in \mathfrak{S} \Rightarrow A \in \mathfrak{S}$. A stack $\mathfrak{G}$ on X is called a *grill* on X if it satisfies the conditions $\varnothing \notin \mathfrak{G}$; $A \cup B \in \mathfrak{G} \Rightarrow A \in \mathfrak{G}$ or $B \in \mathfrak{G}$.

A proximity is usually considered as a relation on X, but since it is assumed to be a symmetric relation it can be taken to be a collection of two element families $[A, B]$. This enables us to make the following definition.

Definition 2.2. A collection Π of families of subsets of X is called a *basic* (or *Čech*) *proximity* on X if it satisfies the requirements:

$$P_0 \colon \mathfrak{A} \in \Pi \Rightarrow |\mathfrak{A}| = 2,$$

$$P_1 \colon |\mathfrak{A}| = 2,\ \bigcap \mathfrak{A} \neq \varnothing \Rightarrow \mathfrak{A} \in \Pi,$$

$$P_2 \colon \Pi(A) \text{ is a grill on } X \text{ for all } A \subset X..$$

The equivalence of this definition to Čech's [4] is established in [19].

In Terwilliger's modification of Ivanova and Ivanov's definition of a contiguity the LO-condition and separatedness are still included. We remove these conditions in defining a basic contiguity.

Definition 2.3. A collection ξ of families of subsets of X is called a *basic contiguity* on X if it satisfies the conditions:

$$C_0 \colon \mathfrak{A} \in \xi \Rightarrow |\mathfrak{A}| < \aleph_0,$$

$$C_1 \colon |\mathfrak{A}| < \aleph_0,\ \bigcap \mathfrak{A} \neq \varnothing \Rightarrow \mathfrak{A} \in \xi,$$

$$C_2 \colon \xi(\mathfrak{A}) \text{ is a grill on } X \text{ for all } \mathfrak{A} \subset \mathfrak{P}(X),$$

$$C_3 \colon \mathfrak{B} \subset \mathfrak{A} \in \xi \Rightarrow \mathfrak{B} \in \xi.$$

For every infinite cardinal number C one can define a *C-contiguity* by

replacing the requirement $|\mathfrak{A}| < \aleph_0$ in C_0 and C_1 by $|\mathfrak{A}| \leq C$.

It is helpful to introduce an operation $\textcircled{\vee}$ as well as a relation $\succ$ on families of subsets of X. We have

$$\mathfrak{A} \textcircled{\vee} \mathfrak{B} = [A \cup B: A \in \mathfrak{A}, B \in \mathfrak{B}]$$

and

$$\mathfrak{B} \succ \mathfrak{A} \Leftrightarrow \forall B \in \mathfrak{B}\ \exists A \in \mathfrak{A} \text{ such that } B \supset A.$$

In terms of this notation we can, following Naimpally [12], define a basic (or Čech) nearness as follows:

Definition 2.4. A collection ν of families of subsets of X shall be called a *basic* (or *Čech*) *nearness* on X if it satisfies:

$$B_1: \bigcap \mathfrak{A} \neq \varnothing \Rightarrow \mathfrak{A} \in \nu,$$

$$B_2: \mathfrak{A} \in \nu \Rightarrow \varnothing \notin \mathfrak{A},$$

$$B_3: \mathfrak{B} \succ \mathfrak{A} \text{ and } \mathfrak{A} \in \nu \Rightarrow \mathfrak{B} \in \nu,$$

$$B_4: \mathfrak{A} \notin \nu,\quad \mathfrak{B} \notin \nu \Rightarrow \mathfrak{A} \textcircled{\vee} \mathfrak{B} \notin \nu.$$

In the sequel it will be convenient to omit the prefixes "basic" or "Čech". Thus a nearness is understood to be a basic nearness and similarly for proximities and contiguities.

With these definitions we are able to prove:

Theorem 2.1. (a) *Let* $|\mathfrak{A}| < \aleph_0$, $|\mathfrak{B}| < \aleph_0$ *and let* ξ *be a contiguity then*

(i) $\mathfrak{B} \succ \mathfrak{A}$ *and* $\mathfrak{A} \in \xi \Rightarrow \mathfrak{B} \in \xi$,

(ii) $\mathfrak{A} \notin \xi, \mathfrak{B} \notin \xi \Rightarrow \mathfrak{A} \textcircled{\vee} \mathfrak{B} \notin \xi$.

(b) $\mathfrak{A} \in \lambda \Rightarrow \mathfrak{A} \subset \lambda(\mathfrak{A})$, *where* λ *may be a proximity, contiguity or nearness.*

(c) *If* ν *is a nearness on* X *then* $\nu(\mathfrak{A})$ *is a grill on* X *for all* $\mathfrak{A} \subset \mathfrak{P}(X)$.

Proof of (a). Let $\mathfrak{B} = [B_1, \ldots, B_m]$. If $\mathfrak{B} \succ \mathfrak{A}$ then for every B_k there exists an $A_k \in \mathfrak{A}$, such that $B_k \supset A_k$. Here $A_k = A_j$, $k \neq j$, is possible. Set $\mathfrak{A}' = [A_1, \ldots, A_m]$, then $\mathfrak{A}' \subset \mathfrak{A}$. Hence by C_3 and the assumption $\mathfrak{A} \in \xi$ we have $\mathfrak{A}' \in \xi$. Set $\mathfrak{A}'_k = [B_1, \ldots, B_k, A_{k+1}, \ldots, A_m]$. Then $\mathfrak{A}'_1 \in \xi$ since from $A_1 \in \xi([A_2, \ldots, A_m])$ it follows by C_2 that $B_1 \in \xi([A_2, \ldots, A_m])$ and hence that $\mathfrak{A}'_1 \in \xi$. Now assume that $\mathfrak{A}'_k \in \xi$. Then $A_{k+1} \in \xi(\mathfrak{A}'_k \sim [A_{k+1}])$ and hence $B_{k+1} \in \xi(\mathfrak{A}'_k \sim [A_{k+1}])$, that is $\mathfrak{A}'_{k+1} \in \xi$. By induction we thus arrive at $\mathfrak{A}'_m = \mathfrak{B} \in \xi$. This establishes (i). We now turn to the proof of (ii).

Set $\mathfrak{A} \circledV \mathfrak{B} = \mathfrak{C}$. Further, let $\mathfrak{A} = [A_1, \ldots, A_n]$, $\mathfrak{B} = [B_1, \ldots, B_m]$ and define $\mathfrak{A}_r = [A_i : i < r]$, $\mathfrak{B}_s = [B_j : j < s]$. Then $\mathfrak{A} = \mathfrak{A}_{n+1}$ and $\mathfrak{B} = \mathfrak{B}_{m+1}$. Assume $\mathfrak{C} \in \xi$. Either $\mathfrak{B}_{m+1} \cup \mathfrak{C} = \mathfrak{B} \cup \mathfrak{C} \in \xi$, in which case $\mathfrak{B} \in \xi$ follows from C_3, or there exists a least k, $1 \leq k \leq m+1$ such that $\mathfrak{B}_k \cup \mathfrak{C} \notin \xi$. Since $\mathfrak{B}_1 = \emptyset$ and $\emptyset \cup \mathfrak{C} = \mathfrak{C} \in \xi$ we must have $k > 1$. Set $r = k - 1$, then $\mathfrak{B}_r \cup \mathfrak{C} \in \xi$. Further, if $B_r \in \xi(\mathfrak{B}_r \cup \mathfrak{C})$ then $[B_r] \cup \mathfrak{B}_r \cup \mathfrak{C} = \mathfrak{B}_{r+1} \cup \mathfrak{C} = \mathfrak{B}_k \cup \mathfrak{C} \in \xi$, which is a contradiction. Hence $\mathfrak{B}_r \cup \mathfrak{C} \in \xi$ and $B_r \notin \xi(\mathfrak{B}_r \cup \mathfrak{C})$. We next observe that since $\mathfrak{A}_1 = \emptyset$ it is true that $\mathfrak{A}_1 \cup \mathfrak{B}_r \cup \mathfrak{C} \in \xi$. Let $1 \leq t \leq m$. Assume that we know that $\mathfrak{A}_t \cup \mathfrak{B}_r \cup \mathfrak{C} \in \xi$. Now $A_t \cup B_r \in \mathfrak{A} \circledV \mathfrak{B} = \mathfrak{C} \subset \mathfrak{A}_t \cup \mathfrak{B}_r \cup \mathfrak{C}$. Hence $A_t \cup B_r \in \xi(\mathfrak{A}_t \cup \mathfrak{B}_r \cup \mathfrak{C})$. Since $B_r \notin \xi(\mathfrak{B}_r \cup \mathfrak{C})$ it is true, a fortiori, that $B_r \notin \xi(\mathfrak{A}_t \cup \mathfrak{B}_r \cup \mathfrak{C})$. This fact, together with C_2, yields $A_t \in \xi(\mathfrak{A}_t \cup \mathfrak{B}_r \cup \mathfrak{C})$ and hence $\mathfrak{A}_{t+1} \cup \mathfrak{B}_r \cup \mathfrak{C} \in \xi$. By induction and recalling that $\mathfrak{A} = \mathfrak{A}_{n+1}$, we conclude that $\mathfrak{A} \cup \mathfrak{B}_r \cup \mathfrak{C} \in \xi$. Using C_3 it then follows that $\mathfrak{A} \in \xi$.

Proof of (b). The result follows from the observation that $A \in \mathfrak{A} \Rightarrow [A] \cup \mathfrak{A} = \mathfrak{A}$.

Proof of (c). Set $\mathfrak{C} = [C] \cup \mathfrak{A}$ and $\mathfrak{D} = [D] \cup \mathfrak{A}$. Then $\mathfrak{C} \circledV \mathfrak{D} \succ [C \cup D] \cup \mathfrak{A}$. Assume $C \cup D \in \nu(\mathfrak{A})$. This is equivalent to $[C \cup D] \cup \mathfrak{A} \in \nu$. Hence by B_3, $\mathfrak{C} \circledV \mathfrak{D} \in \nu$. An application of the contrapositive of B_4 yields $\mathfrak{C} \in \nu$ or $\mathfrak{D} \in \nu$. Thus $C \in \nu(\mathfrak{A})$ or $D \in \nu(\mathfrak{A})$. Now assume that $E \supset F \in \nu(\mathfrak{A})$. Then $[F] \cup \mathfrak{A} \in \nu$ and hence, by B_3, $[E] \cup \mathfrak{A} \in \nu$, that is $E \in \nu(\mathfrak{A})$. Clearly $\emptyset \notin \nu(\mathfrak{A})$ and hence $\nu(\mathfrak{A})$ is a grill.

The analogues of B_3 and B_4 thus are also valid for contiguities. It is also clear that C_2 follows from (a)(i) and (ii). The proof is completely analogous to that for (c). The analogue of C_2 holds for nearness structures. However we are not able to derive B_3 and B_4 from it, since in those axioms we may be dealing with infinite families $\mathfrak{A}$ and $\mathfrak{B}$. The axioms B_3 and B_4 can be derived from the following:

I: $\mathfrak{B} \subset \mathfrak{A}$ and $\mathfrak{A} \in \nu \Rightarrow \mathfrak{B} \in \nu$.

II: Let $\mathfrak{B}$ be well ordered by indexing $\mathfrak{B} = [B_j]$ by means of ordinal numbers and set $\mathfrak{B}_j = [B_i : i < j]$. Then $\mathfrak{B}_i \cup \mathfrak{C} \in \nu \ \forall i < j \Rightarrow \mathfrak{B}_j \cup \mathfrak{C} \in \nu$ provided either $\mathfrak{B} \succ \mathfrak{C}$ or $\mathfrak{C} = \mathfrak{A} \circledV \mathfrak{B}$ for some $\mathfrak{A} \notin \nu$.

The proof resembles the proof of (a) but is by transfinite induction.

Finally, note that if Π is a proximity and $[A, B] \in \Pi$ then $\Pi([A, B]) = [A, B]$. It follows that $\Pi(\mathfrak{A})$ is not always a grill.

Definition 2.5. For a nearness ν on X we define

$$\xi_\nu = [\mathfrak{A}: \mathfrak{A} \in \nu, |\mathfrak{A}| < \aleph_0], \quad \Pi_\nu = [\mathfrak{A}: \mathfrak{A} \in \nu, |\mathfrak{A}| = 2],$$

$$c_\nu(A) = [x: [[x], A] \in \nu].$$

For a contiguity ξ on X we have

$$\Pi_\xi = [\mathfrak{A}: \mathfrak{A} \in \xi, |\mathfrak{A}| = 2], \quad c_\xi(A) = [x: [[x], A] \in \xi].$$

If Π is a proximity on X we define $c_\Pi(A) = [x: [[x], A] \in \Pi]$.

The following results are immediate.

Theorem 2.2. *The family ξ_ν is a contiguity on X. Π_ν and Π_ξ are proximities on X. The functions c_ν, c_ξ, and c_Π are Čech closure operators on X. Finally, $\Pi_\nu = \Pi_{\xi_\nu}$, $c_\nu = c_{\xi_\nu} = c_{\Pi_{\xi_\nu}}$.*

Definition 2.6. A proximity (contiguity, nearness) λ is called a *LO-proximity (contiguity, nearness)* if it satisfies the additional condition: $[c_\lambda(A_i): i \in I] \in \lambda \Rightarrow [A_i: i \in I] \in \lambda$.

Definition 2.7. A proximity (contiguity, nearness) λ is called *separated* if it satisfies the additional condition $[[x], [y]] \in \lambda \Rightarrow x = y$.

A LO-nearness induces a LO-contiguity, which in turn induces a LO-proximity. However a non-LO-nearness may induce a LO-contiguity, and a non-LO-contiguity may induce a LO-proximity. This is illustrated by Examples 2.2 and 2.3 below.

It is well known that the closure operator induced by a LO-proximity (and hence by any LO-structure) is a Kuratowski closure operator.

Definition 2.8. Let λ be a proximity, or contiguity, or nearness on X, then a family $\mathfrak{A} \subset \mathfrak{P}(X)$ is called *λ-compatible* if $\mathfrak{B} \subset \mathfrak{A}$ (and $|\mathfrak{B}| = 2$, or $|\mathfrak{B}| < \aleph_0$, if appropriate) implies $\mathfrak{B} \in \lambda$. The family $\mathfrak{A}$ is called *λ-closed* if $[A] \cup \mathfrak{B} \in \lambda$, for all $\mathfrak{B} \subset \mathfrak{A}$ (and $|\mathfrak{B}| = 1$, or $|\mathfrak{B}| < \aleph_0$, as appropriate), implies $A \in \mathfrak{A}$. If λ is a nearness then a family $\mathfrak{A}$ is λ-closed iff $\lambda(\mathfrak{A}) \subset \mathfrak{A}$.

A λ-compatible grill is called a *λ-clan*. Finally, a λ-closed λ-clan is called a *λ-cluster*.

The next result was proved for separated LO-contiguities by Terwilliger [17]; it is also, in a disguised form, asserted by Herrlich [6] for LO-nearness structures.

Theorem 2.3. *Let λ be a contiguity or a nearness on X. Then every maximal λ-compatible family $\mathfrak{A}$ is a grill and hence a maximal λ-clan on X.*

Proof. Except for the "union property" of $\mathfrak{A}$ the proof is straightforward. We shall consider the case where λ is a contiguity. For nearness

structures the argument is somewhat simpler. Set $\alpha = [\mathfrak{B}: \mathfrak{B} \subset \mathfrak{A}, |\mathfrak{B}| < \aleph_0]$. Let $A \cup B \in \mathfrak{A}$. Then for all $\mathfrak{B} \in \alpha$ we have $[A \cup B] \cup \mathfrak{B} \in \lambda$. Now either $[A] \cup \mathfrak{B} \in \lambda$ for all $\mathfrak{B} \in \alpha$, or $[B] \cup \mathfrak{B} \in \lambda$ for all $\mathfrak{B} \in \alpha$, or there exist families $\mathfrak{B}_1$ and $\mathfrak{B}_2$ in α such that $[A] \cup \mathfrak{B}_1 \notin \lambda$ and $[B] \cup \mathfrak{B}_2 \notin \lambda$. Then by (a)(ii) of Theorem 2.1 $([A] \cup \mathfrak{B}_1) \oslash ([B] \cup \mathfrak{B}_2) = \mathfrak{C} \notin \lambda$. However $\mathfrak{C} \succ [A \cup B] \cup (\mathfrak{B}_1 \cup \mathfrak{B}_2)$ and $\mathfrak{B}_1 \cup \mathfrak{B}_2 \in \alpha$. Thus $[A \cup B] \cup (\mathfrak{B}_1 \cup \mathfrak{B}_2) \in \lambda$. This contradicts (a)(i) of Theorem 2.1. Hence either $[A] \cup \mathfrak{A}$ or $[B] \cup \mathfrak{A}$ is a λ-compatible family. It follows from the maximality of $\mathfrak{A}$ that $A \in \mathfrak{A}$ or $B \in \mathfrak{A}$.

Theorem 2.4. *Let λ be a proximity, or a contiguity, or a nearness. Then every λ-cluster is a maximal λ-clan and a maximal λ-compatible family.*

Theorem 2.5. *Let λ be a contiguity, or a nearness. Then every maximal λ-compatible family is a λ-cluster.*

Proof. Since every maximal λ-compatible family $\mathfrak{A}$ is a λ-clan by Theorem 2.3, it suffices to show that $[A] \cup \mathfrak{B} \in \lambda$ for all $\mathfrak{B} \subset \mathfrak{A}$ with the appropriate cardinality restriction implies $A \in \mathfrak{A}$. This clearly follows from the maximality of $\mathfrak{A}$ as a λ-compatible family.

Theorem 2.6. *If λ is a proximity or contiguity then every λ-compatible family is contained in a maximal λ-compatible family.*

For a contiguity ξ it now follows from Theorems 2.3 and 2.6 that every ξ-compatible family is contained in a maximal ξ-clan. Hence all contiguities are clan generated.

Using Theorem 2.7 it is easy to construct examples of nearness structures ν, where not all ν-compatible families are contained in maximal families.

Example 2.1. Let $X = A \cup B, A \cap B = \emptyset$, $|A| \geq \aleph_0$, $|B| \geq \aleph_0$. Let a closure operator c be defined on X by requiring that X, A, B and all finite sets form a subbase for the closed sets of the space. Define a proximity Π^* on X by: $[C, D] \in \Pi^*$ iff $c(C) \cap c(D) \neq \emptyset$ or both C and D are infinite. Then

$$\mathfrak{M} = [C: C \subset X, |C| \geq \aleph_0]$$
$$= \bigcup [\mathfrak{U}: \text{ is a nonprincipal ultrafilter on } X]$$

is a maximal Π^*-clan. The maximal Π^*-compatible families containing $\mathfrak{M}$ are $\mathfrak{F}_a = [D: a \in D, D \cap B \neq \emptyset] \cup \mathfrak{M}$ and $\mathfrak{F}_b = [D: b \in D, D \cap A \neq \emptyset] \cup \mathfrak{M}$ where $a \in A$, $b \in B$. It is also clear that $\mathfrak{M}$ is not a Π^*-cluster. Even for an EF-proximity Π there may be maximal Π-compatible families which are not clusters. An example can be constructed by considering the three sides of a triangle in the plane.

We can summarize the results obtained above, by considering the following statements:

(A) λ clusters are maximal λ-clans and maximal λ-compatible families,

(B) maximal λ-compatible families are grills,

(C) maximal λ-compatible families are λ-clusters,

(D) λ-compatible families are contained in maximal λ-compatible families,

(E) maximal λ-clans are maximal λ-compatible families.

The following table now gives the desired information:

	A	B	C	D	E
proximity	yes	no Ex. 2.1	no Ex. 2.1	yes	no Ex. 2.1
contiguity	yes	yes	yes	yes	yes
nearness	yes	yes	open	no Th. 2.7	open

Theorem 2.7. *Let* $[\mathfrak{G}_i : i \in I]$ *be a family of grills on* X *with the property that for every* $x \in X$ *there exists an* i *such that* $[x] \in \mathfrak{G}_i$. *Then* ν *defined by* $\mathfrak{A} \in \nu$ *iff* $\mathfrak{A} \subset \mathfrak{G}_i$, *for some* $i \in I$, *is a basic nearness on* X.

Proof. We show that B_4 holds, the other properties are easily seen to be true. If $\mathfrak{A} \notin \nu$ then for every $i \in I$ there exists a set $A_i \in \mathfrak{A}$ such that $A_i \notin \mathfrak{G}_i$. Similarly $\mathfrak{B} \notin \nu$ implies the existence of sets $B_i \in \mathfrak{B}$ such that $B_i \notin \mathfrak{G}_i$. Now assume that $\mathfrak{A} \circledvee \mathfrak{B} \in \nu$. Then there exists a $j \in I$ such that $\mathfrak{A} \circledvee \mathfrak{B} \subset \mathfrak{G}_j$. In particular $A_j \cup B_j$ must be in $\mathfrak{G}_j$ but this is a contradiction since neither A_j nor B_j belongs to $\mathfrak{G}_j$.

The following stronger theorem holds for contiguities.

Theorem 2.8. *Let* $[\mathfrak{G}_i : i \in I]$ *be a family of grills on* X *satisfying the two conditions*

(a) *for every* $x \in X$ *there exists a* $\mathfrak{G}_i$ *such that* $[x] \in \mathfrak{G}_i$,

(b) $i \neq j \Rightarrow \mathfrak{G}_i \not\subset \mathfrak{G}_j$.

Then ξ *defined by* $\mathfrak{A} \in \xi$ *iff* $\mathfrak{A} \subset \mathfrak{G}_i$ *for some* i *and* $|\mathfrak{A}| < \aleph_0$ *is a basic contiguity on* X. *Every contiguity* ξ *on* X *is generated by the family of all its maximal* ξ*-clans.*

Proof. The proof of the first part is completely analogous to the proof of Theorem 2.7. The second part follows from Theorems 2.3 and 2.6.

Theorem 2.8 can be thought of as a representation theorem for contiguity structures since it asserts that all such structures are of the simple type described there.

Theorem 2.9. *Let λ be a proximity, or contiguity, or nearness on X. Let $S \subset X$. Define $\lambda_S = [\mathfrak{A}: \mathfrak{A} \in \lambda \cap \mathfrak{P}(\mathfrak{P}(S))]$, then λ_S is a proximity, or contiguity, or nearness on S. λ_S is called the structure induced by λ on S. Finally, if λ is a* LO*-structure then so is λ_S.*

Definition 2.9. If λ is a proximity, or contiguity, or nearness on X then the pair (X, λ) is called a *proximity space*, or a *contiguity space* or a *nearness space*.

Definition 2.10. A mapping $f: (X, \lambda) \to (Y, \eta)$ is called a *proximity map*, or a *contiguity map*, or a *near map* if $\mathfrak{A} \in \lambda \Rightarrow [f(A): A \in \mathfrak{A}] \in \eta$. The expressions *p-continuous*, *c-continuous* and *n-continuous* are also used.

Definition 2.11. A structure λ will be said to be *larger* that a structure λ' if λ, considered as a collection of families of sets, contains λ'.

We now turn to the discussion of some special nearness and contiguity structures and to some examples.

Definition 2.12. Let ξ be a contiguity on X and let $[\mathfrak{G}_i^\xi : i \in I]$ be the family of all maximal ξ-clans. By $\nu(\xi)$ we shall denote the collection $[\mathfrak{A}: \mathfrak{A} \subset \mathfrak{G}_i^\xi$ for some $i \in I]$.

$[\mathfrak{G}_i^\xi]$ is a family of grills and satisfies the conditions of Theorem 2.7, hence $\nu(\xi)$ is a nearness. Since every $\mathfrak{A} \in \xi$ is contained in a maximal ξ-clan it is true that $\xi_{\nu(\xi)} = \xi$. From this equality it also follows that for every contiguity ξ there is at least one nearness, namely $\nu(\xi)$, which induces it.

Theorem 2.10. *Let ν be a nearness which satisfies the condition: $\mathfrak{A} \notin \nu$ iff $\exists \mathfrak{B} \subset \mathfrak{A}, |\mathfrak{B}| < \aleph_0, \mathfrak{B} \notin \nu$. (That is: ν is a "contigual nearness" as defined by Herrlich* [6].*) Then $\nu = \nu(\xi_\nu)$, where ξ_ν is the contiguity defined in Definition* 2.5. *Moreover, ν is clan generated. Finally, for any contiguity ξ the nearness $\nu(\xi)$ is a contigual nearness.*

Proof. That $\nu = \nu(\xi_\nu)$ can be seen as follows: $\mathfrak{A} \in \nu$ iff $\forall \mathfrak{B} \subset \mathfrak{A}, |\mathfrak{B}| < \aleph_0$, $\mathfrak{B} \in \nu$, iff $\mathfrak{A}$ is a ξ_ν-compatible family, iff $\mathfrak{A} \subset \mathfrak{G}_i$, where $\mathfrak{G}_i$ is a maximal ξ_ν-clan, iff $\mathfrak{A} \in \nu(\xi_\nu)$. Clearly all $\nu(\xi)$ are clan generated. The last assertion can be proved by substituting $\nu(\xi)$ for ν in the first argument and recalling that $\xi_{\nu(\xi)} = \xi$.

Definition 2.13. Let λ be a proximity or a contiguity or a nearness on X. A λ-clan $\mathfrak{G}$ on X will be called a *λ-bunch* iff $b(\mathfrak{G}) = [A: c_\lambda(A) \in \mathfrak{G}] = \mathfrak{G}$.

Theorem 2.11. *If λ is a* LO*-structure on X then every maximal λ-clan on X is a λ-bunch.*

Proof. This is an extension of a theorem proved for proximities by Thron [19].

Now let (X, Π) be given and consider any nearness structure ν such that $\Pi_\nu = \Pi$. Let $[\mathfrak{C}_i^\Pi: i \in I]$ be the collection of all maximal Π-clans. If $\mathfrak{A} \in \nu$ then $\mathfrak{A}$ is ξ_ν-compatible and hence there exists a maximal ξ_ν-compatible family $\mathfrak{H}$, which is a ξ_ν-clan, such that $\mathfrak{A} \subset \mathfrak{H}$. $\mathfrak{H}$ is a Π-clan and hence $\mathfrak{H}$ is contained in one of the maximal Π-clans $\mathfrak{C}_i^\Pi$. It is thus easy to see that there is a largest nearness ν^Π, which induces Π. It is defined as follows: $\nu^\Pi = [\mathfrak{A}: \mathfrak{A} \subset \mathfrak{C}_i^\Pi$ for some $i \in I]$. It is slightly easier to show that $\xi^\Pi = [\mathfrak{A}: |\mathfrak{A}| < \aleph_0, \mathfrak{A} \subset \mathfrak{C}_i^\Pi$ for some $i \in I]$ is the largest contiguity which induces Π.

If $[A, B] \in \Pi$ there are in general several ways (but always at least one) of choosing ultrafilters $\mathfrak{U}_A^B$ and $\mathfrak{U}_B^A$ so that $A \in \mathfrak{U}_A^B$, $B \in \mathfrak{U}_B^A$ and $\mathfrak{G}_{A,B} = \mathfrak{U}_A^B \cup \mathfrak{U}_B^A$ is a Π-clan. (This is shown in [19].) Any collection $[\mathfrak{A}: \mathfrak{A} \subset \mathfrak{G}_{A,B}$ for some $[A, B] \in \Pi]$ is a clan generated nearness. That these nearness structures are "small" in a very real sense follows from the fact that they are generated by grills which are as small as possible, at least for $A \cap B = \emptyset$. If $A \cap B \supset [x]$ one could agree to choose $\mathfrak{U}_A^B = \mathfrak{U}_B^A = \mathfrak{U}(x) = [C: x \in C]$ and thus, in that case, also have $\mathfrak{G}_{A,B} = \mathfrak{U}(x)$ as small as possible. Nevertheless, it will not in general be the case that these near structures are minimal with respect to inducing Π and being clan generated. It is conceivable that certain grills $\mathfrak{G}_{A,B}$ could be deleted from the family of grills defining one of these structures without affecting the proximity induced by the structure. This is always the case if $[A, B] \subset \mathfrak{G}_{C,D}$ for a pair $[C, D] \in \Pi$ and $\mathfrak{G}_{A,B} \neq \mathfrak{G}_{C,D}$. If minimal clan generated nearness structures compatible with a given proximity Π do indeed exist, they must be of the type discussed here.

That for certain proximities Π there does not exist a least near structure compatible with Π has just been shown by Njåstad [15].

The situation becomes much simpler for contiguities. For them we have the theorem:

Theorem 2.12. *Let Π be a given proximity then*

$$\xi_\Pi = [\mathfrak{B}\colon |\mathfrak{B}| < \aleph_0,\ \ \mathfrak{B} = \mathfrak{B}_1 \cup \mathfrak{B}_2,\ \ [\bigcap\mathfrak{B}_1, \bigcap\mathfrak{B}_2] \in \Pi]$$

is the least contiguity compatible with Π. ξ_Π *can also be characterized as*

$$\xi_\Pi = [\mathfrak{B}\colon |\mathfrak{B}| < \aleph_0, \mathfrak{B} \subset \mathfrak{G}_{A,B} \text{ for some } [A, B] \in \Pi].$$

ξ_Π *is independent of the choice of the family* $[\mathfrak{G}_{A,B}\colon [A, B] \in \Pi]$, *provided that for each* $[A, B] \in \Pi$ *there is at least one* $\mathfrak{G}_{C,D}$ *containing* $[A, B]$.

Proof. That the two characterizations define the same collection is easy to check. That the collection is a contiguity compatible with Π can be deduced from the second characterization. Clearly all $\mathfrak{B}$ with $[\bigcap\mathfrak{B}_1, \bigcap\mathfrak{B}_2] \in \Pi$ must be in every ξ compatible with Π. It follows that ξ_Π is the least contiguity compatible with Π.

If Π is a LO-proximity then, by Theorem 2.11, the $\mathfrak{G}_i^\Pi$ are Π bunches so that we have in particular $[c_\Pi(A_j)\colon j \in J] \subset \mathfrak{G}_i^\Pi \Rightarrow [A_j\colon j \in J] \subset \mathfrak{G}_i^\Pi$. It follows that ν^Π and ξ^Π are LO-structures and hence the largest LO-structures inducing Π.

Example 2.2. Let X be the Euclidean plane and Π be the proximity induced by the usual metric on X. Then (X, Π) is a LO-proximity space, but ξ_Π (as defined above) is not a LO-contiguity. To see this let A and B be two disjoint closed sets with $A \in \Pi(B)$. Let

$$c_\Pi(A_1) = c_\Pi(A_2) = A, \qquad A_1 \cap A_2 = \varnothing,$$

$$c_\Pi(B_1) = c_\Pi(B_2) = B, \qquad B_1 \cap B_2 = \varnothing.$$

No union of two ultrafilters can contain the four disjoint sets A_1, A_2, B_1, B_2 hence $[A_1, A_2, B_1, B_2] \notin \xi_\Pi$; however $[c_\Pi(A_1), c_\Pi(A_2), c_\Pi(B_1), c_\Pi(B_2)] \in \xi_\Pi$.

"Small" LO-structures compatible with a given LO-proximity Π can be induced by bunches of the form $b(\mathfrak{G}_{A,B})$. In particular

$$\xi_\Pi^L = [\mathfrak{B}\colon |\mathfrak{B}| < \aleph_0, \mathfrak{B} \subset b(\mathfrak{G}_{A,B}) \text{ for some } [A, B] \in \Pi]$$

is the least LO-contiguity compatible with Π. As before, it can be shown that ξ_Π^L is independent of the choice of the $\mathfrak{G}_{A,B}$. ξ_Π^L can also be described as

$$\xi_\Pi^L = [\mathfrak{B}\colon \mathfrak{B} = [B_1^{(1)}, \ldots, B_n^{(1)}] \cup [B_1^{(2)}, \ldots, B_m^{(2)}],$$

$$[\bigcap[c_\Pi(B_i^{(1)})], \bigcap[c_\Pi(B_i^{(2)})]] \in \Pi].$$

The existence of ξ^{Π} and ξ_{Π}^{L} was known to Terwilliger [17].

Example 2.3. Let X be an infinite set and define $\mathfrak{U}(x) = [A: x \in A]$, $\mathfrak{W} = \bigcup[\mathfrak{U}: \mathfrak{U}$ is a nonprincipal ultrafilter$]$. Define the contiguity ξ on X as follows: $\mathfrak{A} \in \xi$ iff $|\mathfrak{A}| < \aleph_0$ and $\mathfrak{A} \subset \mathfrak{U}(x) \bigcup \mathfrak{W}$ for some $x \in X$. This contiguity induces the minimum T_1-topology on X and is a LO-contiguity. Now define a nearness ν by $\mathfrak{A} \in \nu$ iff $\mathfrak{A} \subset \mathfrak{U}(x) \cup \mathfrak{U}_1 \cup \ldots \cup \mathfrak{U}_k$ where $\mathfrak{U}_1, \ldots, \mathfrak{U}_k$ are arbitrary nonprincipal ultrafilters on X. Clearly, $\xi_\nu = \xi$. ν is not a LO-nearness, since for a LO-nearness every family of infinite sets would have to be near. This is so because in a minimum T_1-space the closures of all these sets would be equal to X.

Example 2.4. Let $X = \bigcup[A_k: k = 1, 2, \ldots]$, where $|A_k|$ is infinite for each k and all A_k are disjoint. Define a contiguity ζ on X by $\mathfrak{A} \in \zeta$ iff $|\mathfrak{A}| < \aleph_0$ and $\mathfrak{A} \subset \mathfrak{U}(x)$, for some x, or $\mathfrak{A} \subset \mathfrak{W}$. ($\mathfrak{U}(x)$ and $\mathfrak{W}$ are as defined in the preceding example.) Then Π_ζ is the proximity in which two sets are close iff they intersect or they are both infinite. This is the largest LO-proximity on X compatible with the discrete topology. Denote by $\mathfrak{G}$ any finite union of nonprincipal ultrafilters and by γ the collection of all of these grills. Then ζ can also be characterized by $\mathfrak{A} \in \zeta$ iff $|\mathfrak{A}| < \aleph_0$ and $\mathfrak{A} \subset \mathfrak{U}(x)$, for some x, or $\mathfrak{A} \subset \mathfrak{G}$ for some $\mathfrak{G} \in \gamma$.

For each k let $\mathfrak{V}_k$ be a nonprincipal ultrafilter containing A_k. Now define a nearness μ on X as follows: $\mathfrak{A} \in \mu$ iff $\mathfrak{A} \subset \mathfrak{U}(x)$ for some $x \in X$, or $\mathfrak{A} \subset \mathfrak{G}$, for some $\mathfrak{G} \in \gamma$, or $\mathfrak{A} \subset \bigcup[\mathfrak{V}_{2k}] = \mathfrak{H}_2$, or $\mathfrak{A} \subset \bigcup[\mathfrak{V}_{2k-1}] = \mathfrak{H}_1$. Clearly $\xi_\mu = \zeta$. For further reference note that $\mathfrak{A}_1 = [A_{2k-1}, k = 1, 2, \ldots] \subset \mathfrak{H}_1$, $\mathfrak{A}_2 = [A_{2k}, k = 1, 2, \ldots] \subset \mathfrak{H}_2$, that $\mathfrak{H}_1 \cup \mathfrak{H}_2 \notin \mu$ but that every finite subset of it is in ζ. We shall use this example in §3 to show that not all Π^ν can be generated by contigual nearness structures $\nu(\xi)$.

The final three results are of importance in §3.

Theorem 2.13. *Let ν be a nearness on X and let $\mathfrak{A} \in \nu$. If $\nu(\mathfrak{A}) \in \nu$ then $\nu(\mathfrak{A})$ is a ν-cluster.*

Theorem 2.14. *If λ is a* LO*-proximity or a* LO*-contiguity or a* LO*-nearness on X then, for every $x \in X$, $\lambda(x)$ is a λ-cluster.*

Theorem 2.15. *Let ν be a nearness on X and let $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ be families of subsets of X such that $\mathfrak{A} \cup \mathfrak{C} \notin \nu$, $\mathfrak{B} \cup \mathfrak{C} \notin \nu$. Then $(\mathfrak{A} \circledvee \mathfrak{B}) \cup \mathfrak{C} \notin \nu$.*

Proof. Define $\mathfrak{C}^* = [A: \exists B \in \mathfrak{C}, A \supset B]$. Then $\mathfrak{A} \cup \mathfrak{C}^* \notin \nu$ and $\mathfrak{B} \cup \mathfrak{C}^* \notin \nu$. It follows from B_4 that

$$\mathfrak{D} = (\mathfrak{A} \cup \mathfrak{C}^*) \oslash (\mathfrak{B} \cup \mathfrak{C}^*) \notin \nu.$$

Now

$$\mathfrak{D} = (\mathfrak{A} \oslash \mathfrak{B}) \cup (\mathfrak{A} \oslash \mathfrak{C}^*) \cup (\mathfrak{B} \oslash \mathfrak{C}^*) \cup (\mathfrak{C}^* \oslash \mathfrak{C}^*).$$

The last three families are all contained in $\mathfrak{C}^*$ since $\mathfrak{C}^*$ is a stack. We thus obtain $\mathfrak{D} \subset (\mathfrak{A} \oslash \mathfrak{B}) \cup \mathfrak{C}^*$ and hence $(\mathfrak{A} \oslash \mathfrak{B}) \cup \mathfrak{C}^* \notin \nu$. In view of condition B_3 this is true iff $(\mathfrak{A} \oslash \mathfrak{B}) \cup \mathfrak{C} \notin \nu$.

3. **Proximities defined in terms of near structures.** We begin this section by reviewing some facts about extensions of topological spaces. We make certain modifications, such as the transition to dual traces, and extend concepts, where feasible, to closure spaces (see Čech [4]).

The triple $(\phi, (Y, d))$ is an *extension* of the closure space (X, c) if (Y, d) is a closure space and ϕ is a homeomorphism from (X, c) to $(\phi(X), d')$, where $d'(A) = d(A) \cap \phi(X)$, $A \subset \phi(X)$, and if $d(\phi(X)) = Y$. That is (X, c) is densely embedded in (Y, d). Two extensions $(\phi, (Y, d))$ and $(\phi^\dagger, (Y^\dagger, d^\dagger))$ are *equivalent* if there exists a homeomorphism ψ from (Y, d) onto $(Y^\dagger, d^\dagger)$ such that $\psi \circ \phi = \phi^\dagger$ on X. If there is no danger of confusion we may sometimes refer to (Y, d) as an extension of (X, c).

For each $y \in Y$ define

$$\tau(y) = \tau(y, (Y, d)) = [A\colon\ A \subset X,\ y \in d(\phi(A))],$$

the *dual trace* of the point y with respect to the extension $(\phi, (Y, d))$. The set $[\tau(y)\colon y \in Y]$ is called the *dual trace system* of the extension.

We note that $\tau(y)$ is a grill on X and hence its dual

$$D(\tau(y)) = [B\colon\ X \sim B \notin \tau(y)] = [B\colon\ B \cap A \neq \emptyset\ \forall A \in \tau(y)]$$

is a filter on X. If d is a Kuratowski closure operator then $\tau(y)$ is a c-grill (a grill $\mathfrak{G}$ is a *c-grill* iff $c(A) \in \mathfrak{G} \Rightarrow A \in \mathfrak{G}$) and $D(\tau(y))$ is an open filter for all $y \in Y$ and $[D(\tau(y))\colon y \in Y]$ is the trace system of the extension $(\phi, (Y, d))$. A simple translation of the usual statement in terms of trace systems gives the following:

Let (X, c) be a T_0-topological space (this insures that $\tau(\phi(x_1)) \neq \tau(\phi(x_2))$ iff $x_1 \neq x_2$, $x_1, x_2 \in X$) and let X^* be a collection of c-grills on X containing all $\Gamma_c(x) = [A\colon x \in c(A)]$. Define

$$A^* = [\mathfrak{G}\colon\ \mathfrak{G} \in X^*,\ A \in \mathfrak{G}],$$

$\phi(x) = \Gamma_c(x)$, and

$$d^*(\alpha) = \bigcap [A^*\colon\ \alpha \subset A^*] \quad \text{for all } \alpha \subset X^*.$$

Then d^* is a Kuratowski closure operator and $(\phi, (X^*, d^*))$ is equivalent to the *principal extension* of (X, c) with respect to the dual trace system X^* (see Thron [18]). The dual trace system of this extension is indeed X^*. More specifically it is true, for every $\mathfrak{G} \in X^*$, that $\tau(\mathfrak{G}) = \mathfrak{G}$.

These (or equivalent) extensions have a number of other names. Banaschewski [1] calls them the *strict extensions* with respect to X^*. Lodato [11] in a more special context obtains the same extensions and Gagrat and Naimpally [5] refer to the topology generated by d^* as the *absorption topology* on X^*. Wagner [21] calls these extensions *filter spaces*.

We next observe that

$$d^*(\phi(A)) = \bigcap[B^*: \phi(A) \subset B^*] = A^* = (c(A))^* = d^*(\phi(c(A))),$$

and that in view of the definition of d^* the family $[A^*: A^* = d^*(\phi(A)), A \subset X]$ forms a base for the closed sets of (X^*, d^*).

An extension $(\phi, (Y, d))$, where the closure operator d is determined by $d(B) = \bigcap[d(\phi(A)): A \subset X, d(\phi(A)) \supset B]$ (this is equivalent to saying that the sets $d(\phi(A))$ form a base for the closed sets of the space) is called by Ivanova and Ivanov [8] a *regular extension*. Thus the principal extension is a regular extension.

Moreover, since the dual trace system (and hence the trace system) of an extension determines the family $[d(\phi(A)): A \subset X]$, and since knowledge of this family determines the dual trace system, there is exactly one (up to equivalent ones) regular extension of a given space for a given dual trace system, namely the principal extension. It is also clear that for any T_0-topological extension $(\phi, (X^*, d))$ of (X, c) with dual trace system X^*, consisting of c-grills, the relation $d(\alpha) \subset d^*(\alpha)$, $\alpha \subset X^*$, holds.

We now turn to the description of a method to define proximity extensions of proximity spaces. Let (X, Π) be a LO-proximity space and let ν be a LO-nearness on X which satisfies $\Pi_\nu = \Pi$. Define

$$X^\nu = [\mathfrak{A}: \mathfrak{A} \text{ is a } \nu\text{-clan}] \supset [\nu(x): x \in X],$$
$$A^\nu = [\mathfrak{A}: \mathfrak{A} \in X^\nu, A \in \mathfrak{A}] \quad \text{for } A \subset X,$$
$$\mathfrak{A}^\nu = [A^\nu: A \in \mathfrak{A}] \quad \text{for } \mathfrak{A} \subset \mathfrak{P}(X).$$

We then let Π^ν be the collection of sets $[\alpha, \beta]$, $\alpha \subset X^\nu$, $\beta \subset X^\nu$ determined as follows: $[\alpha, \beta] \in \Pi^\nu$ iff $(\bigcap\alpha) \cup (\bigcap\beta) \in \nu$. Note that we do not require X^ν to contain all ν-clans. We do however require that it contain all ν-clans of the form $\nu(x)$.

Theorem 3.1. *(X^ν, Π^ν) as defined above is a proximity space.*

Proof. If $\alpha \cap \beta \not\supset [\mathfrak{A}]$ then $(\bigcap \alpha) \cup (\bigcap \beta) \subset \mathfrak{A}$. Since $\mathfrak{A} \in \nu$, $[\alpha, \beta] \in \Pi^\nu$ follows. If $\alpha \not\supset \beta \in \Pi^\nu(\gamma)$ then $(\bigcap \beta) \cup (\bigcap \gamma) \in \nu$ and $\bigcap \alpha \subset \bigcap \beta$. Hence $(\bigcap \alpha) \cup (\bigcap \gamma) \in \nu$ and $\alpha \in \Pi^\nu(\gamma)$. Finally, $\alpha \notin \Pi^\nu(\gamma)$ and $\beta \notin \Pi^\nu(\gamma)$ implies $(\bigcap \alpha) \cup (\bigcap \gamma) \notin \nu$ and $(\bigcap \beta) \cup (\bigcap \gamma) \notin \nu$. An application of Theorem 2.15 yields $((\bigcap \alpha) \oslash (\bigcap \beta)) \cup (\bigcap \gamma) \notin \nu$. Now $\bigcap \alpha$ and $\bigcap \beta$ are stacks. It follows that $(\bigcap \alpha) \oslash (\bigcap \beta) \subset \bigcap(\alpha \cup \beta)$. Thus we obtain $(\bigcap(\alpha \cup \beta)) \cup (\bigcap(\gamma)) \notin \nu$.

Theorem 3.2. *Define* $\phi: (X, \Pi) \to (X^\nu, \Pi^\nu)$ *by* $\phi(x) = \nu(x)$, *for all* $x \in X$, *and let* $c^\nu = c_\Pi \nu$. *Then*

$$\phi(c_\Pi(A)) = A^\nu \cap \phi(X) = c^\nu(A^\nu) \cap \phi(X)$$

for all $A \subset X$, *and*

$$c_\Pi(A) \subset c_\Pi(B) \quad \textit{iff} \quad A^\nu \cap \phi(X) \subset B^\nu \cap \phi(X).$$

Proof. $\nu(x) \in \phi(c_\Pi(A))$ iff $x \in c_\Pi(A)$ that is iff $[[x], A] \in \Pi \subset \nu$, since $\Pi_\nu = \Pi$. Hence $\nu(x) \in \phi(c_\Pi(A))$ iff $A \in \nu(x)$ iff $\nu(x) \in A^\nu \cap \phi(X)$. Next, $\nu(x) \in c^\nu(A^\nu)$ iff $\nu(x) \cup [A] \in \nu$. This is the case, since $\nu(x)$ is a ν-cluster by Theorem 2.14, iff $A \in \nu(x) \in A^\nu \cap \phi(X)$. Though ϕ may not be one-to-one, it is true that $\phi(x) = \phi(x')$ iff $\nu(x) = \nu(x')$ iff x and x' are contained in the same closures.

Theorem 3.3. $\phi: (X, \Pi) \to (X^\nu, \Pi^\nu)$ *is a proximity mapping and* $\phi(X)$ *is dense in* (X^ν, c^ν). *Moreover,* $\phi: (X, c_\Pi) \to (\phi(X), (c^\nu)_{\phi(X)})$ *is a closed mapping. If* Π *is a separated proximity then* ϕ *is one-to-one and provides a proximal embedding of* (X, Π) *into* (X^ν, Π^ν). *Thus* $(\phi, (X^\nu, \Pi^\nu))$ *is a proximity extension of* (X, Π).

Proof. We first observe that for $A \subset X$, $C \in \bigcap \phi(A)$ if $[C] \cup [[a]] \in \nu$ for all $a \in A$ iff $A \subset c_\nu(C)$. Now $[\phi(A), \phi(B)] \notin \Pi^\nu$ iff $(\bigcap \phi(A)) \cup (\bigcap \phi(B)) \notin \nu$. Recalling that ν is a LO-nearness, we then have $[A] \cup [B] \notin \nu$ and thus finally $[A, B] \notin \Pi$. Hence ϕ is a proximity mapping. The remaining assertions of the theorem are easy to verify.

The properties of the space (X^ν, Π^ν) depend very much on the choice of X^ν. Π will continue to be a LO-proximity and ν a LO-nearness on X. Our first result is:

Theorem 3.4. *The proximity* Π^ν *is separated iff for* $\mathfrak{G}_1, \mathfrak{G}_2 \in X^\nu$, $\mathfrak{G}_1 \neq \mathfrak{G}_2$ *it is true that* $\mathfrak{G}_1 \cup \mathfrak{G}_2 \notin \nu$.

Proof. It follows from the definition of Π^ν that $[[\mathfrak{G}_1], [\mathfrak{G}_2]] \in \Pi^\nu$ iff $\mathfrak{G}_1 \cup \mathfrak{G}_2 \in \nu$.

Theorem 3.5. *The trace of the point $\mathfrak{G} \in X^\nu$ with respect to the extension $(\phi, (X^\nu, c^\nu))$ of (X, c_Π) is $\tau(\mathfrak{G}) = \nu(\mathfrak{G})$. If X^ν is such that all $\nu(\mathfrak{G})$ are ν-clans then Π^ν is separated iff $\nu(\mathfrak{G}_1) = \nu(\mathfrak{G}_2) \Rightarrow \mathfrak{G}_1 = \mathfrak{G}_2$. If one thinks of τ as a function from X^ν to $[\nu(\mathfrak{G}): \mathfrak{G} \in X^\nu]$ defined by $\tau(\mathfrak{G}) = \nu(\mathfrak{G})$ then the condition for Π^ν to be separated (provided all $\nu(\mathfrak{G})$ are ν-clans) becomes the one-to-one behavior of τ.*

Proof. $\tau(\mathfrak{G}) = [A: \mathfrak{G} \in c^\nu(\phi(A))] = [A: [A] \cup \mathfrak{G} \in \nu] = \nu(\mathfrak{G})$. If there exist $\mathfrak{G}_1, \mathfrak{G}_2 \in X^\nu$ such that $\mathfrak{G}_1 \neq \mathfrak{G}_2$ and $\nu(\mathfrak{G}_1) = \nu(\mathfrak{G}_2)$, then $\mathfrak{G}_1 \in \nu(\mathfrak{G}_2)$ and, since all $\nu(\mathfrak{G})$ are assumed to be ν-clans, $\mathfrak{G}_1 \cup \mathfrak{G}_2 \in \nu$ so that Π^ν is not separated. If Π^ν is not separated there exist $\mathfrak{G}_1 \neq \mathfrak{G}_2$ such that $\mathfrak{G}_1 \cup \mathfrak{G}_2 \in \nu$. It follows that $\mathfrak{G}_2 \subset \nu(\mathfrak{G}_1)$ which if $\nu(\mathfrak{G}_1) \in \nu$ implies $\nu(\mathfrak{G}_2) \supset \nu(\mathfrak{G}_1)$. By an analogous argument we obtain $\nu(\mathfrak{G}_1) \supset \nu(\mathfrak{G}_2)$ and hence $\nu(\mathfrak{G}_1) = \nu(\mathfrak{G}_2)$.

Theorem 3.6. *If all $\mathfrak{G} \in X^\nu$ are maximal ν-compatible families, then $\mathfrak{G} = \tau(\mathfrak{G})$ for all $\mathfrak{G} \in X^\nu$. Further, Π^ν is a* LO*-proximity on X^ν, the closure operator c^ν is a Kuratowski operator, and $(\phi, (X^\nu, c^\nu))$ is the principal extension of (X, c_Π) with respect to the dual trace system X^ν.*

Proof. If $\mathfrak{G}$ is maximal ν-compatible then $\mathfrak{G} = \nu(\mathfrak{G}) = \tau(\mathfrak{G})$. Let $\alpha \subset X^\nu$ then

$$c^\nu(\alpha) = [\mathfrak{G}: \bigcap \alpha = [A_i: i \in I] \subset \mathfrak{G}] = \bigcap [A_i^\nu].$$

Assume $[\alpha, \beta] \notin \Pi^\nu$ then $\mathfrak{A} = [A_i] = \bigcap \alpha$ and $\mathfrak{B} = [B_j] = \bigcap \beta$ is such that $\mathfrak{A} \cup \mathfrak{B} \notin \nu$. However $c^\nu(\alpha) \subset \bigcap [A_i^\nu]$ and $c^\nu(\beta) \subset \bigcap [B_j^\nu]$ and hence $[c^\nu(\alpha), c^\nu(\beta)] \notin \Pi^\nu$. Thus Π^ν is a LO-proximity and the closure operator induced by it is a Kuratowski operator.

Theorem 3.7. *Let (X^ν, Π^ν) be such that*

(a) $\nu(\mathfrak{G}) \in \nu$, *for all* $\mathfrak{G} \in X^\nu$,

(b) $\mathfrak{G}_1 \neq \mathfrak{G}_2 \Rightarrow \nu(\mathfrak{G}_1) \neq \nu(\mathfrak{G}_2)$, $\mathfrak{G}_1, \mathfrak{G}_2 \in X^\nu$,

(c) c^ν *is a Kuratowski closure operator on* X^ν.

Then $\tau: (X^\nu, c^\nu) \to (X^*, d^*)$, where $\tau(\mathfrak{G}) = \nu(\mathfrak{G})$, is a homeomorphism, and $\tau(\nu(x)) = \nu(x)$. Thus $(\phi, (X^\nu, c^\nu))$ is equivalent to the principal extension with respect to its dual trace system $X^* = [\nu(\mathfrak{G}): \mathfrak{G} \in X^\nu]$. If only (a) *and* (b) *hold one has $\tau(c^\nu(\alpha)) \supset d^*(\tau(\alpha))$.*

Proof. $\nu(\mathfrak{G}) \in \nu$ implies that $\nu(\mathfrak{G})$ are maximal ν-compatible hence the $\nu(\mathfrak{G})$ are also c-grills. The maximality of the $\nu(\mathfrak{G})$ together with the fact that $\nu(\nu(x)) = \nu(x)$ and condition (b) insures that (X, c_Π) is a T_0-space. (As a matter of fact it is a T_1-space since it is generated by a proximity.) This suffices for the existence of the principal extension $(\phi, (X^*, d^*))$ with dual trace system $X^* = [\nu(\mathfrak{G}): \mathfrak{G} \in X^\nu]$. We recall that

$$d^*(\beta) = \bigcap [A^*: \beta \subset A^*], \ \beta \subset X^*,$$

where $A^* = [\nu(\mathfrak{G}): A \in \nu(\mathfrak{G})]$. Thus for $\alpha \subset X^\nu$

$$\begin{aligned} \tau(c^\nu(\alpha)) &= \tau[\mathfrak{G}: \textstyle\bigcap \alpha = [A_i] \subset \nu(\mathfrak{G})] \\ &= [\nu(\mathfrak{G}): \textstyle\bigcap \alpha = [A_i] \subset \nu(\mathfrak{G})] \\ &= \textstyle\bigcap [[\nu(\mathfrak{G}): A_i \in \nu(\mathfrak{G})]: A_i \in \bigcap [\mathfrak{G}: \mathfrak{G} \in \alpha]] \\ &= \textstyle\bigcap [A_i^*: A_i \in \bigcap [\mathfrak{G}: \mathfrak{G} \in \alpha]] \\ &\supset \textstyle\bigcap [A_j^*: A_j \in \bigcap [\nu(\mathfrak{G}): \mathfrak{G} \in \alpha]] \\ &= \textstyle\bigcap [A_j^*: A_j \in \bigcap [\nu(\mathfrak{G}): \nu(\mathfrak{G}) \in \tau(\alpha)]] \\ &= \textstyle\bigcap [A_j^*: \tau(\alpha) \subset A_j^*] = d^*(\tau(\alpha)). \end{aligned}$$

If c^ν is Kuratowski then we must have $\tau(c^\nu(\alpha)) \subset d^*(\tau(\alpha))$. Combining these two inclusion relationships we conclude that τ is a homeomorphism. That $\tau(\nu(x)) = \nu(x)$ follows from the fact, already noted before, that $\nu(\nu(x)) = \nu(x)$.

It would be desirable to have a better sufficient condition for c^ν to be a Kuratowski closure operator than that contained in Theorem 3.6. This however appears to be a difficult problem.

Partly motivated by Lodato's [11] constructions and partly by that employed in the construction of the principal extension we are led to the following definition.

Let (X, Π) be a LO-proximity space and let $X^\dagger$ be a family of Π-bunches on X such that $X^\dagger \supset [\Pi(x): x \in X]$. Define $\phi(x) = \Pi(x)$, $A^\dagger = [\mathfrak{H}: \mathfrak{H} \in X^\dagger, A \in \mathfrak{H}]$, and $d^\dagger(\alpha) = \bigcap [A^\dagger: \alpha \subset A^\dagger]$, $\alpha \subset X^\dagger$. We call $d^\dagger$ the absorption closure operator on $X^\dagger$.

By standard arguments, using the fact that the $\mathfrak{H}$ are grills, one shows that $d^\dagger$ is a Kuratowski closure operator on $X^\dagger$. Moreover, $\phi: (X, c_\Pi) \to (\phi(X), d^\dagger_{\phi(X)})$ is a homeomorphism, since the $\mathfrak{H}$ are c_Π-grills, provided Π is separated.

The fact that the $\mathfrak{H}$ are Π-compatible begins to play an important role only if $X^\dagger$ contains enough $\mathfrak{H}$ so that $[A, B] \in \Pi$ implies the existence of an $\mathfrak{H} \in X^\dagger$ such that $A, B \in \mathfrak{H}$. In this case

$$[A, B] \in \Pi \quad \text{iff} \quad d^\dagger(\phi(A)) \cap d^\dagger(\phi(B)) \neq \emptyset.$$

It is of interest to compare $d^\dagger$ with c^ν in the case $X^\nu = X^\dagger$. We have the following result.

Theorem 3.8. *Let* (X, Π) *be given and let* $X^\nu = X^\dagger$ *be a family of* ν-*bunches (and hence* Π-*bunches). Then* $c^\nu = d^\dagger$ *iff all* $\mathfrak{G} \in X^\nu$ *are maximal* ν-*compatible families.*

Proof. If all $\mathfrak{G}$ are maximal then by Theorem 3.6 $c^\nu = d^\dagger$. If $c^\nu = d^\dagger$, then in particular $c^\nu(A^\nu) = d^\dagger(A^\nu) = d^\dagger(A^\dagger) = A^\dagger = A^\nu$. However, if $\mathfrak{G}$ is not a maximal ν-compatible family then there exists an $A \subset X$ such that $A \notin \mathfrak{G}$ but $[A] \cup \mathfrak{G} \in \nu$. It follows that $\mathfrak{G} \in c^\nu(A^\nu)$ but $\mathfrak{G} \notin A^\nu$ and hence $c^\nu \neq d^\dagger$.

Let ξ be a LO-contiguity on X. For a family X^ξ of ξ-clans one can, in analogy to the definition of Π^ν, define Π^ξ by $[\alpha, \beta] \in \Pi^\xi$ iff $(\bigcap \alpha) \cup (\bigcap \beta)$ is a ξ-compatible family. However, $(X^\xi, \Pi^\xi) = (X^{\nu(\xi)}, \Pi^{\nu(\xi)})$, where $\nu(\xi)$ is the nearness defined in Definition 2.12. First, observe that the $\nu(\xi)$-clans are exactly the ξ-clans, hence any X^ξ is an $X^{\nu(\xi)}$, and conversely. Next, let $[\alpha, \beta] \in \Pi^{\nu(\xi)}$. This is equivalent to $\mathfrak{C} = (\bigcap \alpha) \cup (\bigcap \beta) \in \nu(\xi)$, which is true iff $\mathfrak{C}$ is $\nu(\xi)$-compatible. This is the same as $\mathfrak{C}$ is ξ-compatible, which is the same as $[\alpha, \beta] \in \Pi^\xi$. Thus nothing is lost by not considering separately extension spaces of the form (X^ξ, Π^ξ).

The question remains whether anything is gained by considering extension spaces generated by nearness structures. The answer is in the affirmative as the following example shows:

Let X, ζ, μ, etc. be defined as in Example 2.4. Set

$$X^\mu = X^\zeta = [\mu(x) : x \in X] \cup \gamma \cup [\mathfrak{H}_1, \mathfrak{H}_2].$$

Π^μ and Π^ζ are distinct, since $[[\mathfrak{H}_1], [\mathfrak{H}_2]] \notin \Pi^\mu$ but $[[\mathfrak{H}_1], [\mathfrak{H}_2]] \in \Pi^\zeta$. Π^μ could not be obtained from a smaller contiguity ζ' since for no smaller contiguity will all grills in X^ζ be ζ'-clans.

4. Comparison of extensions (X^ν, Π^ν) with arbitrary extensions. Let (X, Π) be a LO-proximity space and let $(i, (Y, \Pi^*))$, where i is the identity mapping, be an arbitrary proximity extension of (X, Π). The question we propose to investigate in this section is: how close can we come to (Y, Π^*) by a suitable choice of ν and X^ν? It will be convenient to impose on Π^*

a restriction which is slightly weaker then the LO-restriction.

Definition 4.1. Let (X, λ) be a nearness or contiguity or proximity space and let $S \subset X$. Then λ is said to be a LO/*S-structure* iff for all $A_i \subset S$

$$[c_\lambda(A_i): i \in I] \in \lambda \Rightarrow [A_i: i \in I] \in \lambda.$$

Theorem 4.1. *Let $(\phi, (Y, \Pi^*))$ be a proximity extension of a proximity space (X, Π) and let λ^* be a LO/$\phi(X)$ nearness or contiguity or proximity on Y such that $\Pi_{\lambda^*} = \Pi^*$. Define λ on X by $\mathfrak{A} \in \lambda$ iff $\mathfrak{A} \subset \mathfrak{P}(X)$ and $[\phi(A): A \in \mathfrak{A}] \in \lambda^*$. Then every $\tau(y)$ is a λ-clan on X. If in addition c_{λ^*} is a Kuratowski closure operator then the $\tau(y)$ are λ-bunches (and hence a fortiori c_λ-grills).*

Proof. It was already noted in §3 that the $\tau(y)$ are always grills on X. Since all elements of $[c_{\lambda^*}(\phi(A)): A \in \tau(y)]$ have the point y in common, the family $[c_{\lambda^*}(\phi(A)): A \in \tau(y)] \in \lambda^*$. Since $\phi(A) \subset \phi(X)$ and λ^* is LO/$\phi(X)$ it follows that $[\phi(A): A \in \tau(y)] \in \lambda^*$ and hence $\tau(y) \in \lambda$.

The function $\phi: (X, c_\lambda) \to (Y, c_{\lambda^*})$ is continuous; hence $\phi(c_\lambda(A)) \subset c_{\lambda^*}(\phi(A))$. If we now assume that c_{λ^*} is a Kuratowski operator then

$$c_{\lambda^*}(\phi(c_\lambda(A))) \subset c_{\lambda^*}(c_{\lambda^*}(\phi(A))) = c_{\lambda^*}(\phi(A)).$$

Thus $y \in c_{\lambda^*}(\phi(c_\lambda(A)))$ implies $y \in c_{\lambda^*}(\phi(A))$ so that from $c_\lambda(A) \in \tau(y)$ follows $A \in \tau(y)$. Hence if c_{λ^*} is a Kuratowski operator, then each $\tau(y)$ is a λ-bunch.

An immediate consequence of Theorem 4.1 is

Theorem 4.2. *If in $(\phi, (X^\nu, \Pi^\nu))$ Π^ν is a LO/$\phi(X)$-proximity then for all $\mathfrak{S} \in X^\nu$, $\tau(\mathfrak{S}) = \nu(\mathfrak{S})$ are Π-clans.*

Unfortunately, the behavior of the dual trace system of a proximity extension with respect to maximality of its members is not as nice as one might wish. This is illustrated by the following two examples.

Example 4.1. Let $Y = A_1 \supset A_2 \supset A_3 \supset \ldots$ where $\bigcap[A_k: k = 1, 2, \ldots] = \emptyset$. Further let $X \subset Y$ be such that $|(A_k \sim A_{k+1}) \cap X| = \infty$ and that $(A_k \sim A_{k+1}) \cap (Y \sim X) \neq \emptyset$ for all $k \geq 1$. Now let $k(B)$ be the smallest natural number k such that for $|B| = \infty$, $B \sim A_{k(B)}$ is infinite. The proximity Π on Y is then defined as follows: $[A, B] \in \Pi$ iff $c_\Pi(A) \cap C_\Pi(B) \neq \emptyset$, where for finite sets B, $c_\Pi(B) = B$. If $|B| = \infty$ then $c_\Pi(B) = B \cup A_{k(B)-1}$.

Then Π is a separated LO-proximity on Y and $(i, (Y, \Pi))$ has as dual traces, with respect to (X, Π_X), the following: for $y_k \in (A_k \sim A_{k+1}) \cap (Y \sim X)$

$$\tau(y_k) = [B: B \subset X, |B \sim A_{k+1}| = \infty].$$

Thus, clearly, $\tau(y_{k+1}) \supsetneqq \tau(y_k)$ so that no $\tau(y_k)$ is maximal.

Example 4.2. Let $Y = A_1 \cup A_2 \cup A_3$, $A_k \cap A_m = \emptyset$, $k \neq m$, $k = 1, 2, 3$; $m = 1, 2, 3$. $X = A_1' \cup A_2' \cup A_3'$, $A_k' = A_k \cap X$, $|A_k'| = \infty$, $A_k \sim A_k' \neq \emptyset$. Let $\mathfrak{B}_A$ be a nonprincipal ultrafilter on Y containing A and define

$$\mathfrak{D}_C^Z = [B: B \subset Z, |B \cap C| = \infty].$$

It is a grill on Z. Now define c^* on Y by

$$c^*(B) = B \cup \bigcup [A_k: |A_k \cap B| = \infty]$$

and let Π^* be the least separated LO-proximity on Y for which $c_{\Pi^*} = c^*$. Then ν^*, is the least LO/X-nearness on Y which induces Π^*, where ν^* is defined by $\mathfrak{A} \subset \mathfrak{P}(Y)$ is in ν^* iff $\mathfrak{A} \subset \mathfrak{U}(y) \cup \mathfrak{D}^Y_{A_k}$, $y \in A_k$, or $\mathfrak{A} \subset \mathfrak{B}_A \cup \mathfrak{B}_B$, $A \subset Y$, $B \subset Y \sim X$ both infinite, or $\mathfrak{A} \subset \mathfrak{D}^Y_{A_k'} \cup \mathfrak{D}^Y_{A_m'}$. The proximity ν induced by ν^* on X can then be described as follows: $\mathfrak{A} \in \nu$ iff $\mathfrak{A} \subset \mathfrak{P}(X)$ and $\mathfrak{A} \subset \tau(y)$ some $y \in Y$ or $\mathfrak{A} \subset \mathfrak{D}^X_{A_k'} \cup \mathfrak{D}^Y_{A_m'}$. For $y \in A_k \sim A_k'$ we have $\tau(y) = \mathfrak{D}^X_{A_k'}$. These $\tau(y)$ are thus not maximal as ν-clans. We also observe that

$$\nu(\tau(y)) = \mathfrak{D}^X_{A_1'} \cup \mathfrak{D}^X_{A_2'} \cup \mathfrak{D}^X_{A_3'},$$

which is a Π_ν-clan, but is not ν-compatible.

For purposes of comparison it seems desirable to choose X^ν to be the dual trace system of (Y, Π^*). However, since the dual trace system of (X^ν, Π^ν) is $[\nu(\mathfrak{G}): \mathfrak{G} \in X^\nu]$ it becomes clear that our choice can be completely successful only if the $\tau(y) = \mathfrak{G} = \nu(\mathfrak{G})$, that is only if the $\tau(y)$ are all maximal ν-clans. The two preceding examples show that this is not always attainable.

We now show that with the above choice of X, and with a suitable selection of ν, the natural mapping from (Y, Π^*) to (X^ν, Π^ν) is at least a proximity mapping.

Theorem 4.3. *Let $(i, (Y, \Pi^*))$ be an arbitrary proximity extension of the* LO*-proximity space (X, Π). Let ν^* be a* LO/*X-nearness on Y such that $\Pi_{\nu^*} = \Pi^*$. Finally, let $\nu = \nu^* \cap \mathfrak{P}(\mathfrak{P}(X))$. Set $X^\nu = [\tau(y): y \in Y]$ and define $f: (Y, \Pi^*) \to (X^\nu, \Pi^\nu)$ by $f(y) = \tau(y)$ for all $y \in Y$. Then f is a proximity mapping.*

Proof. Clearly $\Pi_\nu = \Pi$ and $r(y)$ is a ν-clan on X (Theorem 4.1). We also have $r(x) = \nu(x)$, for all $x \in X$, and hence $X^\nu = [r(y): y \in Y]$ is a permissible choice.

Let $D \subset X$; then

$$f^{-1}(D^\nu) = [y: y \in Y,\ D \in \Pi^*(y)] = [y: y \in Y,\ y \in c_{\Pi^*}(D)] = c_{\Pi^*}(D).$$

Now let $\alpha, \beta \subset X^\nu$ be such that $[\alpha, \beta] \notin \Pi^\nu$. Then $(\bigcap\alpha) \cup (\bigcap\beta) \notin \nu$. Set $A = f^{-1}(\alpha)$, $B = f^{-1}(\beta)$, $\bigcap\alpha = \mathfrak{A}$, and $\bigcap\beta = \mathfrak{B}$. Then $\mathfrak{A}, \mathfrak{B} \subset \mathfrak{P}(X)$ and

$$A \subset f^{-1}\left(\bigcap[A_i^\nu: A_i \in \mathfrak{A}]\right) = \bigcap[f^{-1}(A_i^\nu): A_i \in \mathfrak{A}] = \bigcap[c_{\Pi^*}(A_i): A_i \in \mathfrak{A}].$$

Similarly $B \subset \bigcap[c_{\Pi^*}(B_j): B_j \in \mathfrak{B}]$. If $[A, B] \in \Pi^*$ then $[A, B] \in \nu^*$ and hence

$$[c_{\Pi^*}(A_i): A_i \in \mathfrak{A}] \cup [c_{\Pi^*}(B_j): B_j \in \mathfrak{B}] \in \nu^*.$$

Since ν^* is a LO/X-nearness and $A_i \subset X$, $B_j \subset X$, it follows that $\mathfrak{A} \cup \mathfrak{B} \in \nu^*$. Since $\mathfrak{A} \cup \mathfrak{B} \subset \mathfrak{P}(X)$ it is also true that $\mathfrak{A} \cup \mathfrak{B} \in \nu$. This is a contradiction and hence $[A, B] \notin \Pi^*$.

The content of the preceding theorem is meaningful only if for any proximity Π^* on Y, which induces a LO-proximity Π on X, there exists a LO/X-nearness ν^* on Y such that $\Pi_{\nu^*} = \Pi^*$. This, though not quite trivial, is indeed the case.

Theorem 4.4. *Let $X \subset Y$ and let (Y, Π^*) be a proximity space such that the proximity induced by Π^* on X is a LO-proximity. Then there exists a LO/X-nearness ν^* such that $\Pi_{\nu^*} = \Pi^*$.*

Proof. For every pair of sets $[C, D] \in \Pi^*$ we determine a pair $\mathfrak{B}_C^D, \mathfrak{B}_D^C$ of ultrafilters on Y such that $C \in \mathfrak{B}_C^D$, $D \in \mathfrak{B}_D^C$, and $\mathfrak{B}_C^D \cup \mathfrak{B}_D^C$ is a Π^*-clan. Next, for any grill $\mathfrak{H}$ on Y define

$$b_X(\mathfrak{H}) = [B: \exists A \subset X,\ B \supset A,\ c_{\Pi^*}(A) \in \mathfrak{H}].$$

It is then easy to verify that $b_X(\mathfrak{H})$ is a grill on Y, and that ν^* defined by $\mathfrak{A} \in \nu^*$ iff

$$\mathfrak{A} \subset \Pi^*(y),\ \text{some } y \in Y,$$

or

$$\mathfrak{A} \subset \mathfrak{B}_C^D \cup \mathfrak{B}_D^C,\ \text{some } [C, D] \in \Pi^*,$$

or

$$\mathfrak{A} \subset b_X(\mathfrak{B}_C^D \cup \mathfrak{B}_D^C),\ \text{some } [C, D] \in \Pi^*,$$

is a LO/X-nearness on Y and satisfies the condition $\Pi_{\nu^*} = \Pi^*$.

Theorem 4.5. *The map f, as defined in Theorem 4.3, is one-to-one iff* (I) $y_1 \neq y_2 \in Y$, *implies the existence of $A \subset X$ such that either* $y_1 \in c_{\Pi^*}(A)$, $y_2 \notin c_{\Pi^*}(A)$ *or* $y_2 \in c_{\Pi^*}(A)$, $y_1 \notin c_{\Pi^*}(A)$, *holds.*

Proof. Condition (I) is exactly what is needed to insure that $\tau(y_1) = \tau(y_2)$ and hence that $f(y) = \tau(y)$ is one-to-one. The condition is mentioned by Ivanov [9] as a possible additional requirement for an extension to be regular. Note that (I) implies but is stronger than the condition that Π^* is separated.

Theorem 4.6. *The function* $f\colon (Y, \Pi^*) \to (X^\nu, \Pi^\nu)$ *provides a proximally isomorphic mapping iff condition* (I) *of Theorem* 4.5 *and*: (J) *For* $A, B \subset Y$, $[A, B] \notin \Pi^* \Rightarrow$ *the existence of collections* $\mathfrak{A}, \mathfrak{B} \subset \mathfrak{P}(X)$ *such that* $A \subset \bigcap [c_{\Pi^*}(A_i)\colon A_i \in \mathfrak{A}]$, $B \subset \bigcap [c_{\Pi^*}(B_j\colon B_j \in \mathfrak{B}]$ *and* $\mathfrak{A} \cup \mathfrak{B} \notin \nu$, *are satisfied. Here f, ν and X^ν are defined as in Theorem* 4.3. *Condition* (J) *implies that all $\tau(y)$ are maximal ν-compatible families.*

Proof. Condition (I) is necessary and sufficient for f^{-1} to exist and (J) is necessary and sufficient for f^{-1} to be a proximity map. The conditions of Theorem 4.3, which are assumed to be satisfied, insure that f is a proximity map. Finally, $A \notin \tau(y)$ implies $[A, [y]] \notin \Pi^*$ and hence by (J), and if we assume that $A \subset X$, it follows that there exists a family $\mathfrak{B} \subset \tau(y) \subset \mathfrak{P}(X)$ such that $[A] \cup \mathfrak{B} \notin \nu$, that is $\tau(y)$ is a maximal ν-compatible family.

Theorem 4.7. *A sufficient condition for all $\tau(y)$ of $(i, (Y, \Pi^*))$ to be maximal ν-compatible families is that for $y \in Y$ and $E \subset X$ $E \notin \tau(y) \Rightarrow \exists N_y$ such that $[E, N_y] \notin \Pi^*$.*

Proof. Since $N_y \cap X = D \neq \emptyset$ we have that $E \notin \tau(y)$ implies the existence of $D \subset X$ such that $[E, D] \notin \Pi$. Hence every $\tau(y)$ is a maximal Π-compatible family and, a fortiori, maximal ν-compatible. A sufficient condition for these conditions to be satisfied is if Π^* is an RH-proximity (see [19]) and hence certainly if Π^* is an EF-proximity.

The conditions appearing in the results above all depend on Π^* as well as on the "position" (a sort of "super density") of X in Y. The stronger the assumptions on Π^* the less important will be the requirements on the "position" of X in Y. Conversely, with relatively weak assumptions on Π^* the requirements on X become critical.

REFERENCES

1. B. Banaschewski, *Extensions of topological spaces*, Canad. Math. Bull. 7 (1964), 1–22. MR 28 #4501.

2. H. L. Bentley, *Extensions of maps on nearness spaces* (preprint).

3. ———, *Nearness spaces and extensions of topological spaces* (preprint).

4. E. Čech, *Topological spaces*, rev. ed., Publ. House Czech. Acad. Sci., Prague; English transl., Wiley, New York, 1966. MR 21 #2962; 35 #2254.

5. M. S. Gagrat and S. A. Naimpally, *Proximity approach to extension problems*, Fund. Math. 71 (1971), 63–76. MR #2653.

6. H. Herrlich, *A concept of nearness*, General Topology and its Appl. (to appear).

7. ———, *Topological structures* (preprint).

8. V. M. Ivanova and A. A. Ivanov, *Contiguity spaces and bicompact extensions of topological spaces*, Izv. Akad. Nauk SSSR Ser. Mat. 23 (1959), 613–634. (Russian) MR 22 #965.

9. A. A. Ivanov, *Regular extensions of topological spaces*, Contributions to Extension Theory of Topological Structures, Berlin, 1969, pp. 133–138.

10. S. Leader, *On clusters in proximity spaces*, Fund. Math. 47 (1959), 205–213. MR 22 #2978.

11. M. W. Lodato, *On topologically induced generalized proximity relations*. II, Pacific J. Math. 17 (1966), 131–135. MR 33 #695.

12. S. A. Naimpally, *Reflective functors via nearness*, Fund. Math. (to appear).

13. S. A. Naimpally and J. H. M. Whitfield, *Not every near family is contained in a clan*, Proc. Amer. Math. Soc. 47 (1975), 237–238.

14. ———, *Proximity extensions* (preprint).

15. O. Njåstad, *A proximity without a smallest compatible nearness* (preprint).

16. Ju. M. Smirnov, *On proximity spaces*, Mat. Sb. 31 (73) (1952), 543–574; English transl., Amer. Math. Soc. Transl. (2) 38 (1964), 5–35. MR 14, 1107.

17. W. L. Terwilliger, *On contiguity spaces*, Ph.D. Dissertation, Washington State University, Pullman, Washington, 1965.

18. W. J. Thron, *Topological structures*, Holt, Rinehart and Winston, New York, 1966. MR 34 #778.

19. ———, *Proximity structures and grills*, Math. Ann. 206 (1973), 35–62.

20. ———, *On a problem of F. Riesz concerning proximity structures*, Proc. Amer. Math. Soc. 40 (1973), 323–326. MR 48 #1179.

21. F. J. Wagner, *Notes on compactification*. I, II, Nederl. Akad. Wetensch. Proc. Ser. A 60 = Indag. Math. 19 (1957), 171–181. MR 19, 436.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF COLORADO, BOULDER, COLORADO 80302

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

AN EMBEDDING THEOREM FOR MATRICES OF COMMUTATIVE CANCELLATIVE SEMIGROUPS

BY

JAMES STREILEIN[1]

ABSTRACT. In this paper it is shown that each semigroup which is a matrix of commutative cancellative semigroups has a "quotient semigroup" which is a completely simple semigroup with abelian maximal subgroups. This result is proved by explicitly constructing the quotient semigroup. The paper also gives necessary and sufficient conditions for a semigroup of the type being considered in the paper to be isomorphic to a Rees matrix semigroup over a commutative cancellative semigroup. Several special cases and examples are also briefly discussed.

The study of the semigroups in the title was initiated by Petrich [7] in connection with commutative separative semigroups. It was conjectured in that paper that matrices of commutative cancellative semigroups can be embedded into Rees matrix semigroups over abelian groups. This paper answers the conjecture affirmatively. We also study the embedding and use it to characterize several special cases of matrices of commutative cancellative semigroups.

0. Preliminaries and summary. We use S to represent a semigroup. If there is a congruence ρ on S for which S/ρ is a rectangular band whose classes are all commutative cancellative semigroups, then we say S is a matrix of commutative cancellative semigroups. Since a rectangular band may be considered as $I \times \Lambda$, the product of a left and right zero semigroup respectively, we will write $S = {}_I\bigcup_\Lambda S_{i\lambda}$ for a matrix of commutative cancellative semigroups, whose classes are the $S_{i\lambda}$. In case the rectangular

Received by the editors March 22, 1974.

AMS (MOS) subject classifications (1970). Primary 20M10, 20M30.

Key words and phrases. Matrix of commutative cancellative semigroups, Rees matrix semigroup, quotient Rees matrix semigroup, quotient group.

(1) This paper contains part of a doctoral dissertation written under the direction of Professor Mario Petrich at Pennsylvania State University.

band above is just Λ, a right zero semigroup, we define a right zero union of commutative cancellative semigroups and write $S = \bigcup_\Lambda S_\lambda$, analogously.

A second concept we will make extensive use of is the Rees matrix semigroup. We denote such a semigroup by $S = \mathfrak{M}(I, G, \Lambda; P)$, where I and Λ are nonempty sets, G is a group, and P maps $\Lambda \times I$ into G. The functional value $P(\lambda, i)$ is denoted by $p_{\lambda i}$. Elements of S are of the form (i, g, λ) with $i \in I$, $g \in G$, $\lambda \in \Lambda$ and multiplication is given by $(i, g, \lambda)(j, h, \mu) = (i, gp_{\lambda j}h, \mu)$. We call S the Rees matrix semigroup over the group G with sandwich matrix P. It is a well-known theorem in semigroup theory that a semigroup is completely simple if and only if it is isomorphic to a Rees matrix semigroup over some group. A completely simple semigroup is a simple semigroup which contains an idempotent e which has the property that if f is another idempotent for which $f = ef = fe$, we must have $e = f$. Any other concepts not defined in the text may be found in Petrich [6] or Clifford and Preston [3].

The main result of §1 is that matrices of commutative cancellative semigroups are precisely subsemigroups of Rees matrix semigroups over abelian groups. We do this by constructing a special Rees matrix semigroup, called the quotient Rees matrix semigroup, into which a given matrix of commutative cancellative semigroups can be embedded. A characterization of a special type of matrix of commutative cancellative semigroups is given.

§2 contains the justification for calling the particular Rees matrix semigroups over abelian groups constructed in §1 a quotient Rees matrix semigroup. This is given in a theorem which says that its quotient Rees matrix semigroup is the smallest into which a matrix of commutative cancellative semigroups can be embedded. There are also several other results which give further information about the nature of the embedding.

In §3 we use the results already obtained in §§1 and 2 to characterize Rees matrix semigroups over commutative cancellative semigroups, which generalize the notion of a Rees matrix semigroup over a group. We also consider a restricted family of Rees matrix semigroups over commutative cancellative semigroups.

§4 contains a short discussion of several examples. These include free contents, prime quasi-uniserial semigroups and $\mathfrak{N}$-semigroups.

1. The embedding. We start with several definitions which have been used to characterize matrices of commutative cancellative semigroups

and a lemma which is probably known.

If for any a, b, $c \in S$ we have $abc = bac$, then we call S left commutative.

Lemma. *If $S = \bigcup_\Lambda S_\lambda$ is a right zero union of commutative cancellative semigroups, then S is left commutative.*

Proof. Let $a, b, c \in S$, so that $a \in S_\lambda$, $b \in S_\mu$, and $c \in S_\eta$ for some $\lambda, \mu, \eta \in \Lambda$. Since S_η is a commutative semigroup, we compute $(abc)(bc) = (bc)(abc) = b(c)(abc) = b(abc)c = (ba)(bc)c = (bac)(bc)$. Therefore we have $abc = bac$, by cancellation in S_η, as required.

We need the following two definitions before we can present our first theorem. We define S to be *weakly cancellative* if for $a, b, x \in S$, $ax = bx$ and $xa = xb$ implies $a = b$. A semigroup S is *conditionally* commutative if for $a, b \in S$ with $ab = ba$, then for any $c \in S$ we have $acb = bca$.

Theorem 1. *The following conditions on a semigroup S are equivalent:*

(i) *S is a matrix of commutative cancellative semigroups.*

(ii) *S is weakly cancellative and conditionally commutative.*

(iii) *S can be embedded in a Rees matrix semigroup over an abelian group.*

Proof. As mentioned in the introduction, Petrich [7] has proved the equivalence of conditions (i) and (ii). Therefore we will start with $S = {}_I\bigcup_\Lambda S_{i\lambda}$, a matrix of commutative cancellative semigroups, which is weakly cancellative and conditionally commutative. We will construct a Rees matrix semigroup, $Q_a(S)$, over an abelian group into which S can be embedded.

To start the construction, we fix $1 \in I$, $1 \in \Lambda$, an element $a \in S_{11}$ and let G be the quotient group over S_{11}, written in the natural way as quotients of elements in S_{11}. We also define a mapping P from $\Lambda \times I$ into G by

$$P_{\lambda i} = \frac{a^2sta^2}{asa^2ta} \quad \text{for some } s \in \bigcup_I S_{j\lambda},\ t \in \bigcup_\Lambda S_{i\mu}.$$

To show that P is single valued we choose another $u \in \bigcup_I S_{j\lambda}$ and $v \in \bigcup_\Lambda S_{i\mu}$ and will show

$$\frac{asta}{asa^2ta} = \frac{auva}{aua^2va}.$$

We obtain the following string of equalities:

$$(asta)(aua)(ava) = (asta)(ava)(aua) = (as)(ta^2)(va^2)(ua)$$

$$= (as)(va^2)(ta^2)(ua) \quad \text{(by commutativity in } S_{i1})$$

$$= (asva)(ata)(aua) = (asva)(aua)(ata) = (asva)(au)(a^2ta)$$

$$= (au)(asva)(a^2ta) \quad \left(\text{by left commutativity in } \bigcup_{\Lambda} S_{1\mu}\right)$$

$$= (au)(as)va^3ta = (as)(au)va^3ta \quad \text{(by commutativity in } S_{1\lambda})$$

$$= (as)(auva)(a^2ta) = (auva)(as)(a^2ta) \quad \left(\text{by left commutativity in } \bigcup_{\Lambda} S_{1\mu}\right)$$

$$= (auva)(asa)(ata).$$

Thus we have established that P is single valued.

Therefore $Q_a(S) = \mathfrak{M}(I, G, \Lambda; P)$ is a Rees matrix semigroup over an abelian group. Define a function ϕ_a on S by:

$$\phi_a(b) = (i, aba/a^2, \lambda) \quad \text{for } b \in S_{i\lambda}.$$

It is immediate that ϕ_a is a function from S into $Q_a(S)$. Let $b, c \in S$ with $b \in S_{i\lambda}$ and $c \in S_{j\mu}$. For these elements,

$$\phi_a(b)\phi_a(c) = \left(i, \frac{aba}{a^2}, \lambda\right)\left(j, \frac{aca}{a^2}, \mu\right) = \left(i, \frac{aba}{a^2} p_{\lambda j} \frac{aca}{a^2}, \mu\right)$$

$$= \left(i, \frac{aba}{a^2} \cdot \frac{a^2bca^2}{(aba)(aca)} \cdot \frac{aca}{a^2}, \mu\right) = \left(i, \frac{abca}{a^2}, \mu\right) = \phi_a(bc)$$

using the definition of $p_{\lambda j}$. Hence ϕ_a is a homomorphism.

Let $b, c \in S$ be such that $\phi_a(b) = \phi_a(c)$. Then $(i, aba/a^2, \lambda) = (j, aca/a^2, \mu)$, implying $i = j$, $\lambda = \mu$ and $aba/a^2 = aca/a^2$. Thus $b, c \in S_{i\lambda}$ and $bc = cb$. This implies $bac = cab$ by conditional commutativity. Multiplying by a, we have $abac = acab$ and $baca = caba$. Using $aba = aca$ we obtain $abac = abab$ and $baba = caba$ giving $ac = ab$ and $ba = ca$ by cancellation in the respective subsemigroups. These equalities imply $c = b$ by weak cancellation. Thus ϕ_a is one-to-one and is actually an embedding.

Conversely, it is immediate that a subsemigroup of a Rees matrix semigroup over an abelian group is a matrix of commutative cancellative semigroups.

We call the Rees matrix semigroup $Q_a(S)$, constructed in the above theorem, the *quotient Rees matrix semigroup* for S. We note that $p_{\lambda i}$ in the theorem is the identity if $\lambda = 1$ or $i = 1$.

We next use Theorem 1 to give a new proof of a part of the following theorem from Petrich [7]. This theorem characterizes medial, weakly cancellative semigroups. A *medial semigroup* is one which satisfies the identity, $abcd = acbd$. We will also have occasion to use *square commutativity* which means that we always have $(ab)^2 = a^2b^2$ for $a, b \in S$. Finally a *rectangular abelian group* is the direct product of a rectangular band and an abelian group.

Theorem 2. *The following conditions on a semigroup S are equivalent:*

(i) *S is medial and weakly cancellative.*

(ii) *S is a matrix of cancellative semigroups and is square commutative.*

(iii) *S is embeddable into a rectangular abelian group.*

(iv) *S is a subdirect product of a rectangular band and a commutative cancellative semigroup.*

Proof. We give only the proof of "(ii) implies (iii)" and refer the reader to Petrich [7] for the remainder. Let S be a matrix of cancellative semigroups, which is also square commutative. Say $S = {}_I\bigcup_\Lambda S_{i\lambda}$. If $a, b \in S_{i\lambda}$, then $a^2b^2 = (ab)^2$. This implies $ab = ba$ by cancellation in $S_{i\lambda}$. Hence S is a matrix of commutative cancellative semigroups. By Theorem 1 we know S can be embedded into the Rees matrix semigroup $Q_a(S)$ over an abelian group G. Let $(i, a, \lambda), (j, b, \mu) \in S$. By hypothesis $(i, a, \lambda)^2(j, b, \mu)^2 = ((i, a, \lambda)(j, b, \mu))^2$. Hence $(i, ap_{\lambda i}ap_{\lambda j}bp_{\mu j}b, \mu) = (i, ap_{\lambda j}bp_{\mu i}ap_{\lambda j}b, \mu)$, so that $p_{\lambda i}p_{\mu j} = p_{\lambda j}p_{\mu i}$, which implies $p_{\mu j}p_{\lambda j}^{-1}p_{\lambda i} = p_{\mu i}$. This is exactly the requirement given in Petrich [6, IV. 3.3], that a Rees matrix semigroup is the direct product of a rectangular band and a group.

If $xa = xb$ for $a, b, x \in S$ implies $a = b$, then S is *left cancellative.* Analogously to a rectangular abelian group, a *right abelian group* is the direct product of a right zero semigroup and an abelian group. These concepts are used in the following corollary of Theorem 2, which is proved in Petrich [7].

Corollary. *The following conditions on a semigroup S are equivalent:*

(i) *S is left commutative and left cancellative.*

(ii) *S is embeddable into a right abelian group.*

(iii) *S is a subdirect product of a commutative cancellative semigroup and a right zero semigroup.*

(iv) *S is a right zero union of commutative cancellative semigroups.*

2. Quotient Rees matrix semigroups. We immediately give the theorem which justifies our earlier definition. As a corollary we will have the earlier known result for right zero unions of commutative cancellative semigroups. We then develop further properties of the embedding ϕ constructed in Theorem 1.

Theorem 3. *Let $S = {}_I\bigcup_\Lambda S_{i\lambda}$ be a matrix of commutative cancellative semigroups, and ϕ_a be the embedding of S into $Q_a(S)$ given in the proof of Theorem 1. If θ is a homomorphism of S into T, a completely simple semigroup, then there exists ψ a unique homomorphism of $Q_a(S)$ into T which makes the following diagram commutative:*

$$\begin{tikzcd} S \arrow[r, "\phi_a"] \arrow[dr, "\theta"'] & Q_a(S) \arrow[d, dashed, "\psi"] \\ & T \end{tikzcd}$$

Proof. We will let S_{11} be the subsemigroup of S used to construct ϕ_a and $Q_a(S) = M(I, G, \Lambda; P)$ as in the proof of Theorem 1. As we noted after Theorem 1, P has all entries in the row and the column containing P_{11} equal to the identity.

By the Rees theorem for completely simple semigroups $T \cong M(I', H, \Lambda'; Q)$ for some group H. Since θ is a homomorphism it must take elements that commute to elements that commute. Hence θ induces mappings of I into I' and Λ into Λ'. We will denote these mappings by primes so that if $b \in S_{i\lambda}$, then $\theta(b) = [i', b', \lambda'] \in T$, where we are using square brackets to distinguish more readily those of T from those of $Q_a(S)$. As an additional simplification, we will require that all entries of Q in the row and column containing $q_{1'1'}$ be identity elements, which can be done following [3, 3.4].

We next define a mapping $\omega\colon S_{11} \to H$ by $\theta(b) = [1', \omega(b), 1']$ $(b \in S_{11})$. Since S_{11} generates its quotient group G, we can extend ω to all of G by $\omega(cb^{-1}) = \omega(c)(\omega(b))^{-1}$. It is easy to verify that ω is a homomorphism on G.

Using ω, we define the mapping $\psi\colon M(I, G, \Lambda; P) \to M(I', H, \Lambda'; Q)$ by $\psi(i, b, \lambda) = [i', \omega(b), \lambda']$ $(b \in G)$. We will show that ψ is the required mapping.

Let $(i, b, \lambda), (1, c, \mu) \in Q_a(S)$. Then

$$\psi((1, b, \lambda)(1, c, \mu)) = \psi(1, bc, \mu) = [1', \omega(bc), \mu']$$
$$= [1', \omega(b)\omega(c), \mu'] = [1', \omega(b)q_{\lambda' 1'}\omega(c), \mu']$$
$$= [1', \omega(b), \lambda'][1', \omega(c), \mu'] = \psi(1, b, \lambda)\psi(1, c, \mu).$$

Therefore it is clear that ψ is a homomorphism when restricted to $\bigcup_\Lambda S_{i\lambda}$. Similarly it can be shown that ψ restricted to $\bigcup_I S_{i1}$ is a homomorphism. We also note that it is immediate from the definition of ω and ψ, that we have $\theta(b) = \psi\phi_q(b)$ for all b in S_{11}, so the diagram commutes on S_{11}.

For b in S_{11} and c in $S_{1\lambda}$, we have cb in S_{11}. Thus

$$\theta(c)\theta(b) = \theta(cb) = \psi\phi_a(cb) = \psi\phi_a(c)\psi\phi_a(b) = \psi\phi_a(c)\theta(b).$$

Since $\theta(c)$ and $\psi\phi(c)$ must be in the same subgroup of T, we must have $\theta(c) = \psi\phi(c)$. Hence the diagram commutes on $\bigcup_I S_{i1}$ and similarly it will also commute on $\bigcup_\Lambda S_{1\lambda}$.

We now let c be in $S_{1\lambda}$ and d be in S_{i1}, so that $\phi(c) = (1', c', \lambda')$ and $\phi(d) = (i', d', 1')$ for some c', d' in G. Therefore

$$[1', \omega(c')q_{\lambda' i'}\omega(d'), 1'] = [1', \omega(c'), \lambda'][i', \omega(d'), 1']$$
$$= \theta(c)\theta(d) = \theta(cd) = \psi\phi_a(cd)$$
$$= \psi((1, c', \lambda)(i, d', 1)) = \psi(1, c'p_{\lambda i}d', 1)$$
$$= [1', \omega(c'p_{\lambda i}d'), 1] = [1', \omega(c')\omega(p_{\lambda i})\omega(d'), 1'],$$

which implies that $q_{\lambda' i'} = \omega(p_{\lambda i})$.

We are finally ready to show that ψ is a homomorphism. Let (i, b, λ), $(j, c, \mu) \in Q_a(S)$. Then

$$\psi((i, b, \lambda)(j, c, \mu)) = \psi(i, bp_{\lambda j}c, \mu)$$
$$= [i', \omega(bp_{\lambda j}c), \mu'] = [i', \omega(b)\omega(p_{\lambda j})\omega(c), \mu']$$
$$= [1', \omega(b)q_{\lambda' j'}\omega(c), \mu']$$
$$= [i, \omega(b), \lambda'][j', \omega(c), \mu'].$$

Thus ψ is a homomorphism as required.

We still need to show that the diagram commutes for any b in S. We already have this by definition for b in S_{11} and have shown this for all b in $\bigcup_\Lambda S_{1\lambda}$ and $\bigcup_I S_{i1}$. Therefore we only have to check commutativity for an element c in $S_{i\lambda}$. We let $b \in S_{11}$, so that bcb is in S_{11}. Then $\theta(bcb) =$

$\psi\phi_a(bcb)$. Hence $\theta(b)\theta(c)\theta(b) = \theta(bcb) = \psi\phi_a(b)\psi\phi_a(c)\psi\phi_a(b) = \theta(b)\psi\phi_a(c)\theta(b)$. Since $\theta(c)$ and $\psi\phi_a(c)$ are in the same subgroup, it follows that $\theta(c) = \psi\phi(c)$. Thus we have shown that the diagram commutes. It is also clear from the proof that any other map ψ' which makes the diagram commutative must take the same action as ψ and thus is the same function and the theorem is proved.

The following corollary for right zero unions of commutative cancellative semigroups is due to Dickinson [4].

Corollary. *Let $S = \bigcup_\Lambda S_\lambda$ be a right zero union of commutative cancellative semigroups, and ϕ be the embedding of S into $Q_a(S)$ given in Theorem 1. If θ is an embedding of S into T, a right abelian group, then there exists ψ, a unique embedding of $Q_a(S)$ into T which makes the following diagram commutative:*

$$\begin{tikzcd} S \arrow[r, "\phi"] \arrow[dr, "\theta"'] & Q_a(S) \arrow[d, dashed, "\psi"] \\ & T \end{tikzcd}$$

The next proposition shows that quotient Rees matrix semigroups are, up to isomorphism, not dependent upon the choice of the element a used in the construction in Theorem 1.

Theorem 4. *Let $S = {}_I\bigcup_\Lambda S_{i\lambda}$ be a matrix of commutative cancellative semigroups, and let $a \in S_{i\lambda}$ and $b \in S_{j\mu}$. If we let ϕ_a, ϕ_b be the embeddings of S in $Q_a(S)$, $Q_b(S)$ as constructed in the proof of Theorem 1, using a, b respectively, then there exists an isomorphism ψ of $Q_a(S)$ onto $Q_b(S)$ which makes the following diagram commutative:*

$$\begin{tikzcd} S \arrow[r, "\phi_a"] \arrow[dr, "\phi_b"'] & Q_a(S) \arrow[d, dashed, "\psi"] \\ & Q_b(S) \end{tikzcd}$$

Proof. This follows immediately from Theorem 3. We have unique homomorphisms $\psi_{a,b}: Q_a(S) \to Q_b(S)$ and $\psi_{b,a}: Q_b(S) \to Q_a(S)$ such that $\psi_a\psi_{a,b} = \phi_b$ and $\phi_b\psi_{b,a} = \phi_a$ but then $\phi_a\psi_{a,b}\psi_{b,a} = \phi_a$. Thus $\psi_{a,b}\psi_{b,a}$

is the identity map on $Q_a(S)$ and similarly $\psi_{b,a}\psi_{a,b}$ is the identity map on $Q_b(S)$. Hence $\psi_{a,b}$ and $\psi_{b,a}$ are inverse isomorphisms.

Corollary. *For $S = {}_I\bigcup_\Lambda S_{i\lambda}$ a matrix of commutative cancellative semigroups, the image of each $S_{i\lambda}$ under any ϕ in the proof of Theorem 1 generates the group into which it is embedded.*

This is just one of the results needed in the proof of the proposition.

The following corollary could also be derived from the already mentioned work of Dickinson [4].

Corollary. *For $S = {}_I\bigcup_\Lambda S_{i\lambda}$, a matrix of commutative cancellative semigroups, all the $S_{i\lambda}$ have isomorphic quotient groups.*

3. Rees compositions. In this section we generalize the construction of Rees matrix semigroups to any semigroup and we characterize those semigroups obtained by using commutative cancellative semigroups in this way. The special case of the direct product of a rectangular band and a commutative cancellative semigroup is also studied.

We need to introduce several standard concepts. A *left translation* λ is a function, written on the left, of S to S which satisfies $\lambda(xy) = \lambda(x)y$ for $x, y \in S$. A *right translation* ρ is defined similarly when written on the right. A left translation λ and a right translation ρ are *linked* if $x(\lambda y) = (x\rho)y$ for $x, y \in S$. The *translation hull* of a semigroup, denoted by $\Omega(S)$, is the set of pairs of linked left and right translations, (λ, ρ), considered as bitranslations. If $(\lambda, \rho), (\lambda', \rho') \in \Omega(S)$ then multiplication defined by $(\lambda, \rho)(\lambda', \rho') = (\lambda\lambda', \rho\rho') \in \Omega(S)$ makes $\Omega(S)$ a semigroup. It is also clear that $(\iota, \iota) \in \Omega(S)$, where ι is the identity function on S written on the proper side, is the identity for $\Omega(S)$. Hence one can consider the group of units of $\Omega(S)$. A left translation λ and a right translation ρ are *permutable* if $(\lambda x)\rho = \lambda(x\rho)$ for all $x \in S$. A set of bitranslations T is *permutable* if for any (λ, ρ), $(\lambda', \rho') \in T$, we have that λ and ρ' are permutable.

We extend the definition of Rees matrix semigroups to $T = \mathfrak{M}(I, S, \Lambda; P)$, where I, Λ are any nonempty sets and S is any semigroup. However P maps $\Lambda \times I$ into a permutable subset of the group of units of $\Omega(S)$. It can be verified that this definition produces a semigroup when $(i, a, \lambda), (j, b, \mu) \in T$ multiply as $(i, a, \lambda)(j, b, \mu) = (i, ap_{\lambda j}b, \mu)$, where $(ap_{\lambda j})b = a(p_{\lambda j}b) = ap_{\lambda j}b$. We call $\mathfrak{M}(I, S, \Lambda; P)$ a *Rees matrix semigroup* over the semigroup S.

Since we are mainly concerned with matrices of commutative cancellative semigroups, we will consider here only Rees matrix semigroups over

commutative cancellative semigroups. It has been shown by Hall [5] and Dickinson [4] that $\Omega(S)$ for a commutative cancellative semigroup consists of exactly those elements g of the quotient group, say G, of S such that $gS \subseteq S$, i.e. the idealizer of S in G. It is also immediate that, because of commutativity, all elements $m \in \Omega(S)$ are permutable. Hence in the case of a Rees matrix semigroup over a commutative cancellative semigroup we only require that $p_{\lambda i}$ be a member of the group of units of the idealizer of S in its quotient group. We now present a theorem which characterizes Rees matrix semigroups over commutative cancellative semigroups.

Theorem 5. *Let* $S = {}_I\bigcup_\Lambda S_{i\lambda}$ *be a matrix of commutative cancellative semigroups. The following statements are equivalent*:

(i) *For all* $a, b \in S$, *there exists* $c, d \in S$ *such that* $cba = bac$, $ba = ca$, $dab = abd$, $ab = ad$.

(ii) *For all* $a \in S$, *if* $a \in S_{i\lambda}$ *then* $aS_{j\mu} = aS_{i\mu}$ *and* $S_{j\mu}a = S_{j\lambda}a$.

(iii) S *is isomorphic to a Rees matrix semigroup over any* $S_{i\lambda}$.

Proof. (i) *implies* (ii). Let $a \in S_{i\lambda}$ and $b \in S_{j\mu}$. By the hypothesis of (i) we have an element d such that $dab = abd$ and $ab = ad$. Thus $d \in S_{i\mu}$ and we have $aS_{j\mu} \subseteq aS_{i\mu}$. Therefore $baS_{i\mu} \subseteq baS_{j\mu}$ since $ba \in S_{j\lambda}$. Since we are in a subsemigroup of a Rees matrix semigroup over a group, we have $aS_i \subseteq aS_j$ and the first equality in (ii) holds. The second equality in (ii) follows similarly.

(ii) *implies* (iii). Fix $a \in S_{i\lambda}$ and construct the embedding of Theorem 1. We claim that the image of S in $\mathfrak{M}(I, G, \Lambda; P) = T$, where G is the quotient group of $S_{i\lambda}$, is a subset of T of the form $I \times S_{i\lambda} \times \Lambda$.

To show this we observe from the proof of Theorem 1 that this is equivalent to showing that $\phi(S_{j\mu}) = \{j\} \times S_{i\lambda} \times \{\mu\}$ or equivalently $aS_{j\mu}a = a^2S_{i\lambda}$. Using the hypothesis in (ii),

$$(aS_{j\mu})a = a(S_{i\mu}a) = aS_{i\lambda}a = a^2S_{i\lambda}$$

proving that the image is as claimed.

It only remains to show that the $p_{\mu j}$ are in the group of units of the idealizer of $S_{i\lambda}$ in G. We already know this if $j = i$ or $\mu = \lambda$ since all such $p_{\mu j}$ are the identity as seen from the proof of Theorem 1. Let (j, b, μ) and (k, c, λ) be in the image of S in T. Then $(j, b, \mu)(k, c, \lambda) = (j, d, \lambda)(k, c, \lambda)$ for some (j, d, λ) in the image of S in T by (ii). Hence $bp_{\mu k}c = dp_{\lambda k}c = dc$ and thus $bp_{\mu k} = d$. This shows that $p_{\mu k}$ is in the idealizer of $S_{i\lambda}$ in G.

By (ii) for each $d \in S_{i\lambda}$ there exists a $b \in S_{i\lambda}$ such that $(j, b, \mu)(k, c, \lambda) = (j, d, \lambda)(k, c, \lambda)$ which implies $bp_{\mu k} = d$, showing that $p_{\mu k}$ takes $S_{i\lambda}$ onto $S_{i\lambda}$. Thus T is a Rees matrix semigroup over $S_{i\lambda}$.

(iii) *implies* (i). Let $S = \mathfrak{M}(I, T, \Lambda; P)$ be a Rees matrix semigroup over the commutative cancellative semigroup T. Then for $a, b \in S$ we have $a = (i, a', \lambda)$ and $b = (j, b', \mu)$. It is immediately verified that $c = (j, b'p_{\mu i}p_{\lambda i}^{-1}, \lambda)$ and $d = (i, b'p_{\lambda j}p_{\lambda i}^{-1}, \mu)$ are the elements needed in (i).

Note. Professor Petrich has suggested that the conditions on the sandwich matrix, P, can be relaxed in the case of Rees matrix semigroups over commutative cancellative semigroups and he has characterized such semigroups.

We also mention that a result entirely similar to that for direct products of rectangular bands and groups as mentioned in the proof of Theorem 2 can be proved for Rees matrix semigroups over any semigroup. We use this result in the next theorem to characterize direct products of rectangular bands and commutative cancellative semigroups.

Theorem 6. *A semigroup S is isomorphic to the direct product of a rectangular band and a commutative cancellative semigroup if and only if S is weakly cancellative, medial and for any $a, b \in S$ there exist $c, d \in S$ for which $bca^2 = ca^2c$ and $a^2db = da^2d$.*

Proof. If S is isomorphic to $T \times B$ where T is a commutative cancellative semigroup and B is a rectangular band, then by Theorem 2, S is weakly cancellative and medial. We represent B as $I \times \Lambda$ with I a left and Λ a right zero semigroup, respectively. Let $(a, (i, \lambda)), (b, (j, \mu)) \in T \times B$. It is immediately checked that the elements $(b, (j, \lambda))$ and $(b, (i, \mu))$ satisfy the requirements in the statement of this theorem for c and d, respectively.

Conversely, assume S is weakly cancellative, medial and satisfies the requirements on elements in the theorem. By Theorem 2, S is a matrix of commutative cancellative semigroups, say $S = {}_I\bigcup_\Lambda S_{i\lambda}$. Let $a, b \in S$ so that by hypothesis there exist $c, d \in S$ with $bca^2 = ca^2c$ and $a^2db = da^2d$. If $a \in S_{i\lambda}$, $b \in S_{j\mu}$ then $c \in S_{j\lambda}$ and $d \in S_{i\mu}$. We have $bca^2 = baca$ by mediality and $aca = a^2c$ by right commutativity in $\bigcup_I S_{k\lambda}$ from the corollary after Theorem 2. Thus $ba^2c = ca^2c$ and since by Theorem 1 we are in a subsemigroup of a Rees matrix semigroup over an abelian group we have $ba = ca$. Hence by Theorem 5, S is a Rees matrix semigroup over any of the commutative cancellative semigroups $S_{i\lambda}$.

We now let (i, a, λ), (j, b, μ), (k, c, μ) and $(l, d, \gamma) \in S$. By mediality $(i, a, \lambda)(j, b, \mu)(k, c, \mu)(l, d, \gamma) = (i, a, \lambda)(k, c, \mu)(j, b, \mu)(l, d, \gamma)$. This implies that $p_{\lambda j} p_{\mu k} p_{\mu l} = p_{\lambda k} p_{\mu j} p_{\mu l}$, or $p_{\lambda j} p_{\mu j}^{-1} p_{\mu k} = p_{\lambda k}$ using commutativity and cancellation in the quotient group of the $S_{i\lambda}$ for which we do the embedding in Theorem 1. Hence, by the result referred to in Theorem 2, S is actually isomorphic to the direct product of the commutative cancellative semigroup $S_{i\lambda}$ and the rectangular band $I \times \Lambda$.

Corollary. *For a semigroup S the following are equivalent:*

(i) *S is a Rees matrix semigroup of commutative, cancellative semigroups with $|I| = 1$.*

(ii) *S is a left commutative, left cancellative semigroup and for $a, b \in S$ there exists a $c \in S$ such that $baca = ca^2c$.*

(iii) *S is left commutative, left cancellative and $Sa \subseteq aS$ for all $a \in S$.*

(iv) *S is isomorphic to the direct product of a commutative, cancellative semigroup and a right zero semigroup.*

The equivalence of (i), (iii) and (iv) can be found in Petrich [7].

4. Examples. We discuss briefly free contents, prime quasi-uniserial semigroups and Rees matrix semigroups over $\mathfrak{N}$-semigroups.

As defined by Tamura [10], the *free content* on two generators, denoted by $C(a, b)$, is the subsemigroup of $F(a, b)$, the free semigroup on the two generators a and b, which consists of all words that contain both a and b at least once.

It has been shown by Shafer [8] that any countable semigroup can be embedded in $C(a, b)$.

Shafer [8] denotes by λ the congruence on $F(a, b)$ generated by the identities $a = a^2$ and $b = b^2$. He has shown that $C(a, b)/\lambda$ is a matrix of infinite cyclic semigroups.

As a second example we mention prime quasi-uniserial semigroups as defined by Behrens [1], [2]. Let I be any set, G be the infinite cyclic group generated by ω, C be the subsemigroup of G consisting of $\{\omega^s | s = 0, 1, \ldots\}$, and π be a function from $I \times I$ to the nonnegative integers, $\{0, 1, 2, \ldots\}$, which satisfies the following conditions. If we denote $\pi(i, j)$ by (ij), π must satisfy:

1. $(ii) = 0$,
2. $(ij) + (jk) \geq (ik)$,
3. $(kj) + (ji) > 0$, $i \neq j$.

The set $S = I \times G \times I$ is a *prime quasi-uniserial semigroup* when we define multiplication by $(h, \omega^s, i)(j, \omega^t, k) = (h, \omega^{s+t+(ij)}, k)$. It is easy to see that S is a matrix of commutative cancellative semigroups. The conditions on π are also equivalent to the conditions that all $e_i = (i, \omega^0, i)$, $i \in I$, are idempotent and that $T = \bigcup_{i,j} Ce_ie_j$ is a subsemigroup of S containing no further idempotents, where $\omega^s(h, \omega^t, i) = (h, \omega^{t+s}, i)$. It is easily checked that T is a matrix of commutative cancellative semigroups. Behrens uses such semigroups in the study of prime, arithmetic rings with identity.

In fact both of the above examples can be considered as the more restrictive case of matrices of $\mathfrak{N}$-semigroups. An $\mathfrak{N}$*-semigroup* is an archimedean commutative cancellative semigroup without idempotents. Tamura [9] has constructed all $\mathfrak{N}$-semigroups as pairs (G, I), where G is an abelian group and I maps $G \times G$ into N, the nonnegative integers, and satisfies the following conditions:

(i) $I(\alpha, \beta) + I(\alpha\beta, \gamma) = I(\alpha, \beta\gamma) + I(\beta, \gamma)$ $(\alpha, \beta, \gamma \in G)$,

(ii) $I(\alpha, \beta) = I(\beta, \alpha)$ $(\alpha, \beta \in G)$,

(iii) $I(\epsilon, \epsilon) = 1$ where ϵ is the identity of G,

(iv) for each $\alpha \in G$ there exists $m > 0$ such that $I(\alpha^m, \alpha) > 0$.

The multiplication on $S = N \times G$ defined by $(m, \alpha)(n, \beta) = (m + n + I(\alpha, \beta), \alpha\beta)$ makes S an $\mathfrak{N}$-semigroup.

Hall [5] has characterized the group of units of the idealizer of S in its quotient group, which we denote by $\Sigma(S)$. The characterization is that

$$\Sigma(S) = \{[0, g] \mid g \in G,\ I(g, h) > 0, \text{ and } I(g^{-1}, h) > 0 \\ \text{for all } h \in G,\ I(g, g^{-1}) = 1\},$$

where $[0, g]$ is a function on S defined by $[0, g](n, h) = (n + I(g, h) - 1, gh)$ using the Tamura representation (G, I) given above.

If G is any abelian group then $I\colon G \times G \to \{1\}$ satisfies the above four conditions. For $S = N \times G$, $\Sigma(S) = \{[0, g] \mid g \in G\}$. We can use these facts to construct many Rees matrix semigroups over the $\mathfrak{N}$-semigroup S.

As another less trivial example, let $G = \{e, a, a^2, a^3\}$ be the cyclic group of order 4. If we define I by $I(e, a^i) = 1$ for $i = 1, 2, 3, 4$, $I(a, a) = 0$, $I(a, a^2) = 1$, $I(a, a^3) = 2$, $I(a^2, a^2) = 3$, $I(a^3, a^2) = 3$ and $I(a^3, a^3) = 2$, then I satisfies the above conditions and $\Sigma(S) = \{[0, e], [0, a^2]\}$.

REFERENCES

1. E. Behrens, *Ring theory*, Academic Press, New York, 1972.

2. E. Behrens, *The arithmetic of the quasi-uniserial semigroups without zero*, Canad. J. Math. 23 (1971), 507–516. MR 44 #2677.

3. A. H. Clifford and G. B. Preston, *The algebraic theory of semigroups*. Vol. I, Math. Surveys, no. 7, Amer. Math. Soc., Providence, R. I., 1961. MR 24 #A2627.

4. R. P. Dickinson, Jr., *On right-zero unions of commutative semigroups*, Pacific J. Math. 41 (1972), 355–364. MR 46 #5488.

5. R. E. Hall, *The translational hull of an $\mathfrak{N}$-semigroup*, Pacific J. Math. 41 (1972), 379–389. MR 46 #5495.

6. M. Petrich, *Introduction to semigroups*, Merrill, Columbus, Ohio, 1973.

7. ———, *Normal bands of commutative cancellative semigroups*, Duke Math. J. 40 (1973), 17–32. MR 47 #381.

8. J. Shafer, *Homomorphisms and subdirect products of free contents* (unpublished).

9. T. Tamura, *Commutative nonpotent archimedian semigroup with cancellation law*. I, J. Gakugei Tokushima Univ. 8 (1957), 5–11. MR 20 #3224.

10. ———, *The study of closets and free contents related to the semi-lattice decomposition of semigroups*, Semigroups (Proc. Sympos., Wayne State Univ., Detroit, Mich., 1968), Academic Press, New York, 1969, pp. 221–260. MR 46 #5504.

DEPARTMENT OF MATHEMATICS, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PENNSYLVANIA 16802

THE ARMY MATERIAL SYSTEMS ANALYSIS AGENCY, ABERDEEN PROVING GROUNDS, ABERDEEN, MARYLAND 21005

Current address: R.D. 5 Box 218, Elkton, Maryland 21921

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

POLAR SETS AND PALM MEASURES IN THE THEORY OF FLOWS(1)

BY

DONALD GEMAN AND JOSEPH HOROWITZ

ABSTRACT. Given a flow (θ_t), t real, over a probability space Ω, we prove that certain measures on Ω (viewed as the state space of the flow) decompose uniquely into a Palm measure Q which charges no "polar set" and a measure supported by a polar set. Considering the continuous and discrete parts of the additive functional corresponding to Q, we find that Q further decomposes into a measure charging no "semipolar set" and a measure supported by one. As a consequence, Palm measures are exactly those which neglect sets which the flow neglects, and polar sets are exactly those neglected by every Palm measure. Finally, we characterize various properties, such as predictability and continuity, of an additive functional in terms of its Palm measure. These results further illuminate the role played by supermartingales in the theory of flows, as pointed by J. de Sam Lazaro and P. A. Meyer.

0. Introduction. Let $(\Omega, \mathcal{F}_t^0, P, \theta_t)$, $t \in \mathbf{R}$ (the real line), be a filtered dynamical system (all terminology will be explained below). In §1 we prove that a finite measure Q on $\mathcal{F}_{0-}^0$ which is "progressively absolutely continuous" decomposes uniquely into the sum of two measures $Q = P_\alpha^- + \mu$, where P_α^- is the restriction to $\mathcal{F}_{0-}^0$ of the Palm measure P_α of a predictable additive functional α, and μ is supported by a "polar" set in $\mathcal{F}_{0-}^0$. Decomposing α into its continuous and discrete parts, say α_c and α_d, we will see (in §2) that P_α^- splits into a measure P_c^- which charges no "semipolar" set and a measure P_d^- which is carried by a semipolar, but charges no polar, set. Thus we have a decomposition

$$Q = P_c^- + P_d^- + \mu \tag{1}$$

analogous to that of a measure on the state space of a Markov process [1,

Received by the editors April 11, 1974.

AMS (MOS) subject classifications (1970). Primary 28A65, 60G10.

Key words and phrases. Flow, filtration, dynamical system, Palm measure, additive functional, polar set, semipolar set.

(1) This work was partially supported by National Science Foundation grant GP 34485 A1.

p. 283]; this is not entirely surprising in view of the Markovian nature of the flow θ_t: $(\Omega, \mathcal{F}_t^0) \to (\Omega, \mathcal{F}_0^0)$ (see [9]). The decomposition (1) requires the Doob-Meyer decomposition of supermartingales and is related to Föllmer's [3] correspondence between supermartingales and certain measures on $\mathbf{R}_+ \times \Omega$. As a corollary, we find that a finite measure Q on $\bigvee_{t \in \mathbf{R}} \mathcal{F}_t^0$ is a Palm measure iff it charges no polar set. The section concludes with a characterization of polar sets, and several applications of these ideas, particularly to local times.

In §2 we characterize various properties of a given additive functional a, such as well-measurability, predictability, and continuity, in terms of its Palm measure. Papangelou [14] has recently given some results on stationary point processes (which we construe as additive functionals which increase only by unit jumps) of the type we are considering. (Similar questions for Markov additive functionals have been treated by Revuz [16].) In the present article we will generalize several of Papangelou's results and obtain flow theory analogues of some of the Markovian ones. Some of our material will be recognized as a specialization of results in the "general theory of processes" [2], with more detail made possible by additional structure. In fact, our intention throughout is to solidify further the bridge built between the general theory and flow theory by J. de Sam Lazaro and P. A. Meyer [8], [9].

The remainder of this section is devoted to an explanation of the terminology and background material. Our notation is largely that of [1, Chapter 0], with this exception: if $(E, \mathcal{E})$ is a measurable space, we write (ambiguously) $f \in (\mathcal{E})$ to mean that f is an $\mathcal{E}$-measurable function on E, the range being clear from context; $f \in (\mathcal{E})_+$ indicates the range is $\mathbf{R}_+ = [0, \infty)$.

A *flow* $\theta = (\theta_t)$, $t \in \mathbf{R}$, on a probability space $(\Omega, \mathcal{F}^0, P)$ is a one-parameter group (under composition) of bimeasurable, measure-preserving bijections θ_t: $\Omega \to \Omega$ such that θ_0 = identity and the mapping $(t, \omega) \to \theta_t(\omega)$ is $\mathcal{B} \otimes \mathcal{F}^0/\mathcal{F}^0$-measurable. We further assume the existence of a *filtration*, i.e. an increasing family of σ-fields $\{\mathcal{F}_t^0\}$, $t \in \mathbf{R}$, on Ω whose generated σ-field $\bigvee_{t \in \mathbf{R}} \mathcal{F}_t^0$ is $\mathcal{F}^0$, and which is compatible with the flow θ in that $\theta_t^{-1}\mathcal{F}_s^0 = \mathcal{F}_{s+t}^0$, $s, t \in \mathbf{R}$. As usual we write $\mathcal{F}_{t+}^0 = \bigcap_{s>t} \mathcal{F}_s^0$, $\mathcal{F}_{t-}^0 = \bigvee_{s<t} \mathcal{F}_s^0$. Each $\mathcal{F}_t^0$ (and thus $\mathcal{F}^0$) is assumed separable. The P-completion of $\mathcal{F}^0$ is denoted $\mathcal{F}$, and then $\mathcal{F}_t$ is obtained by adjoining to $\mathcal{F}_t^0$ all sets in $\mathcal{F}$ of measure zero. The family $\{\mathcal{F}_t\}$ is then right-continuous [2] and compatible with θ, but θ: $(t, \omega) \to \theta_t(\omega)$ need not be $\mathcal{B} \otimes \mathcal{F}/\mathcal{F}$-measurable. The entity $(\Omega, \mathcal{F}_t^0, P, \theta_t)$ is a *filtered dynamical system*. Concepts from the general theory of processes, such as predictability, are in reference to the family $\{\mathcal{F}_t\}$ unless otherwise indicated.

An *additive functional* (AF) is a real-valued process $\alpha = \alpha(t, \omega)$ (or $\alpha_t(\omega)$), $t \in \mathbf{R}$, $\omega \in \Omega$, such that (i) $\alpha(0) = 0$; (ii) almost every path is right-continuous, nondecreasing; (iii) for each $s, t \in \mathbf{R}$ there is a set $N_{st} \in \mathcal{F}$ of measure zero such that

$$\alpha(t + s, \omega) = \alpha(t, \omega) + \alpha(s, \theta_t\omega) \tag{2}$$

for $\omega \notin N_{st}$. By (ii) we may consider α as a measure on $\mathcal{B}$. Notice that α need not be adapted; sometimes we will use the phrase (due to Getoor and Sharpe) "raw additive functional" (RAF) to emphasize this point. We call α *adapted* if $\alpha(t) \in (\mathcal{F}_t)$ for $t \geq 0$, and this implies $\alpha(t) \in \mathcal{F}_0$ if $t \leq 0$. As shown in [9], there exists an AF $\bar{\alpha}$ indistinguishable from α (i.e. such that $\bar{\alpha}(t, \omega) = \alpha(t, \omega)$ for all $t \in \mathbf{R}$ a.s.) which is *perfect* in that the set N_{st} in (iii) may be chosen independently of s, t, $\mathcal{F}^0$-measurable, and such that (iv) $\bar{\alpha}(\pm\infty, \omega) = \pm\infty$ or $\bar{\alpha}(t, \omega) \equiv 0$ for every $\omega \in \Omega$.

Given an AF α, its *Palm measure* is

$$P_\alpha(A) = E\int_0^1 I_A \circ \theta_t \, d\alpha_t \quad \left(= E\int_0^\infty e^{-t} I_A \circ \theta_t \, d\alpha_t\right), \qquad A \in \mathcal{F}^0, \tag{3}$$

where I_A is the indicator of A. Palm measures arise naturally in the study of "flows under a function" [7], local times [4], "time-changes" of flows [17], point processes ("Palm-Khinhin formulae" etc.), and level crossings ("horizontal-window" probabilities). They are exactly the measures which neglect sets in Ω which the flow neglects (Theorem (10)).

Finally, we will need these facts (see [5], [9]). P_α is always σ-finite and finite iff $E\alpha(1) < \infty$, in which case α is called *integrable* and $E\alpha(t) = tE\alpha(1)$. Two AF's α, β are indistinguishable iff their Palm measures are identical. (In particular, $P_\alpha = P_{\bar{\alpha}}$.) In addition, if α and β are both adapted (resp. predictable), then it is enough for the Palm measures to agree on $\mathcal{F}^0_{0+}$ (resp. $\mathcal{F}^0_{0-}$).

1. Decomposition theorems and characterizations of Palm measures. A function $\xi \in (\mathcal{F}^0)$ for which $\xi \circ \theta_t(\omega) \to \xi(\omega)$ as $t \downarrow 0$ for all $\omega \in \Omega$ will be called *translation continuous*. It is shown in [9] that there exists another filtration $\{\mathcal{G}^0_t\}$ such that $\mathcal{G}^0_t \subset \mathcal{F}^0_{t-}$, $\mathcal{G}_t = \mathcal{F}_t$ for all $t \in \mathbf{R}$, where $\{\mathcal{G}_t\}$ is the completed family corresponding to $\{\mathcal{G}^0_t\}$ (see §0), and $\mathcal{G}^0 = \bigvee_{t \in \mathbf{R}} \mathcal{G}^0_t$ is generated by the translation continuous functions. Moreover, the mapping $(t, \omega) \to \theta_t(\omega)$ is $\mathcal{B} \otimes \mathcal{G}^0/\mathcal{G}^0$-measurable. Since the two filtrations differ by sets of measure zero only, there is no essential loss of generality in assuming that $\mathcal{F}^0$ is itself generated by the translation continuous functions. For reasons (in addition to those above) which will soon be apparent, *we assume from now on*

(4) (I) $\mathcal{F}^0$ is generated by the translation continuous functions,
(II) $(\Omega, \mathcal{F}^0)$ is a Blackwell space.

The meaning of (II) is: (i) $\mathcal{F}^0$ is separable, and (ii) for every real-valued $\xi \in (\mathcal{F}^0)$ and $A \in \mathcal{F}^0$, the image $\xi(A)$ is analytic in $\mathbf{R}$. The basic fact we will require is this: let $(E, \mathfrak{E})$ be a Blackwell space and $\mathcal{G}$ a separable sub σ-field of $\mathfrak{E}$. A function $f \in (\mathfrak{E})$ which is constant on the atoms of $\mathcal{G}$ is then $\mathcal{G}$-measurable. (See [13] and [9, Appendix Chapter I].)

We hasten to add that many of the results below do not depend on (I), though they are more complicated without it, and that most of the standard spaces arising in flow theory satisfy (I) and (II). Two examples are the function spaces $\mathfrak{C}$ and $\mathfrak{D}$ of all continuous functions (resp. right-continuous functions having left limits) from $\mathbf{R}$ to $\mathbf{R}$; for $f \in \mathfrak{C}$ (or $\mathfrak{D}$) we let $\theta_t f(s) = f(s+t)$, $X_s f = f(s)$, and $\mathcal{F}_t^0 = \sigma\{X_s\colon s \leq t\}$. Another example, arising in connection with point processes, is the space $\mathfrak{W}$ of all locally finite (i.e. having no finite accumlation point) nonempty subsets of $\mathbf{R}$. Here, for $w \in \mathfrak{W}$, we let $\theta_t w = w - t$ and take $\mathfrak{W}_t^0$ as the σ-field generated by the functions $N(A, w) =$ cardinality of $A \cap w$ for Borel sets $A \subset (-\infty, t]$. These examples will be discussed below.

Let $Z \in (\mathcal{F}_0)_+$ be such that $Z_t = e^{-t} Z \circ \theta_t$, $t \in \mathbf{R}_+$, is a supermartingale relative to $\{\mathcal{F}_t\}$; Z is then called *excessive*. We need the following result, which may be derived from [8] or [9].

(5) **Theorem.** *Let Z be excessive. Then there exists an excessive $Z^0 = Z$ a.s. for which $Z^0 \in (\mathcal{F}^0_{0+})_+$ and the mapping $t \to Z^0 \circ \theta_t(\omega)$ is right-continuous and has left limits at every $t \in \mathbf{R}$, for almost every $\omega \in \Omega$.*

In particular, let $Z = (Z_t)$, $t \in \mathbf{R}_+$, be a potential [13] such that $Z_t = e^{-t} Z_0 \circ \theta_t$ a.s. for each t (Z is "almost homogeneous"). Then there exists a *homogeneous potential* $Z_t^0 = e^{-t} Z^0 \circ \theta_t$, with Z^0 as described in (5), such that Z_t^0 and Z_t are indistinguishable.

Let $\mathcal{B}_+$ denote the Borel σ-field in $\mathbf{R}_+$, and define, for any process $u \in (\mathcal{B}_+ \otimes \mathcal{F}^0)$, two new processes, $\theta^+ u$ and $\theta^- u$, by

$$\theta^+ u(s, \omega) = u(s, \theta_s \omega), \qquad \theta^- u(s, \omega) = u(s, \theta_{-s} \omega).$$

These are measurable, and θ^+ and θ^- are obviously inverses of one another. Now define $\mathcal{P}^0$ to be the σ-field on $\mathbf{R}_+ \times \Omega$ generated by all sets of the form $[t, \infty) \times A$, $t \in \mathbf{R}_+$, $A \in \mathcal{F}^0_{t-}$ (equivalently: all sets of the form $\{0\} \times A$, $A \in \mathcal{F}^0_{0-}$, and $(t, \infty) \times A$, $t \in \mathbf{R}_+$, $A \in \mathcal{F}^0_{t-}$). This is similar to the usual predictable σ-field [2], but more appropriate in the present context.

(6) **Lemma.** $\mathcal{P}^0 = \theta^+(\mathcal{B}_+ \otimes \mathcal{F}^0_{0-})$, *i.e.* $u \in (\mathcal{P}^0)$ *iff* $\theta^- u \in (\mathcal{B}_+ \otimes \mathcal{F}^0_{0-})$.

First note that $(\mathbf{R}_+ \times \Omega, \mathcal{B}_+ \otimes \mathcal{F}^0)$ is a Blackwell space, and $\mathcal{B}_+ \otimes \mathcal{F}^0_{0-}$ is a separable sub σ-field of $\mathcal{B}_+ \otimes \mathcal{F}^0$. Consider a generator of $\mathcal{P}^0$, say $[t, \infty) \times A$, with $A \in \mathcal{F}^0_{t-}$, and let $u(s, \omega) = I_{[t,\infty)}(s) I_A(\omega)$. We wish to show $\theta^- u$ is constant on the atoms of $\mathcal{B}_+ \otimes \mathcal{F}^0_{0-}$. Suppose $(s, \omega), (s', \omega')$ are contained in such an atom. Then, for every $B \in \mathcal{B}_+$, $C \in \mathcal{F}^0_{0-}$, we have $I_B(s) I_C(\omega) = I_B(s') I_C(\omega')$, so $s = s'$ and ω, ω' lie in the same atom of $\mathcal{F}^0_{0-}$. Hence $\theta^- u(s, \omega) = I_{[t,\infty)}(s) I_A(\theta_{-s}\omega)$ and $\theta^- u(s', \omega') = I_{[t,\infty)}(s) I_A(\theta_{-s}\omega')$ both vanish if $s \le t$, and are equal if $s \ge t$, because then $\theta_s A \in \mathcal{F}^0_{(t-s)-} \subset \mathcal{F}^0_{0-}$. We have shown $\mathcal{P}^0 \subset \theta^+(\mathcal{B}_+ \otimes \mathcal{F}^0_{0-})$. Next, let $v \in (\mathcal{B}_+ \otimes \mathcal{F}^0_{0-})$ be of the form $v(s, \omega) = I_B(s) I_C(\omega)$, $B \in \mathcal{B}_+$, $C \in \mathcal{F}^0_{0-}$. Noting that $\mathcal{P}^0$ is a separable sub σ-field of $\mathcal{B}_+ \otimes \mathcal{F}^0$, we will show that $\theta^+ v$ is constant on atoms of $\mathcal{P}^0$. Let $(s, \omega), (s', \omega')$ be in an atom of $\mathcal{P}^0$. Then $s = s'$, and $I_A(\omega) = I_A(\omega')$ for every $A \in \mathcal{F}^0_{s-}$. Thus $v(s, \theta_s\omega) = I_B(s) I_C(\theta_s\omega) = I_B(s') I_C(\theta_s\omega)$ since $\theta_s^{-1} C \in \mathcal{F}^0_{s-}$. The proof is complete.

(7) **Corollary.** (a) *If* $t \in \mathbf{R}_+$ *and* $A \in \mathcal{F}^0_{t+}$, *then* $(t, \infty) \times A \in \mathcal{P}^0$.
(b) *If* $\xi \in (\mathcal{F}^0_{0-})$, *then the process* $\xi \circ \theta = (\xi \circ \theta_t(\omega))$ *is* $\mathcal{P}^0$*-measurable.*

Part (a) results from $(t, \infty) \times A = \bigcup_{n=1}^\infty (t + n^{-1}, \infty) \times A$; (b) is trivial.

To each measure Q on $\mathcal{F}^0_{0-}$ we now associate a measure $\tilde{Q}$ on $\mathcal{P}^0$ as follows. Writing $\tilde{Q}(u)$ for $\int u \, d\tilde{Q}$,

$$\tilde{Q}(u) = \int_0^\infty e^{-s} \int_\Omega \theta^- u(s, \omega) Q(d\omega)\, ds, \qquad u \in (\mathcal{P}^0)_+. \tag{8}$$

We further write $\tilde{Q}_t(A) = \tilde{Q}[(t, \infty) \times A]$, $A \in \mathcal{F}^0_{t+}$, which is possible by (7)(a), and call $\tilde{Q}$ (or Q) *progressively absolutely continuous* (relative to P) if, for each $t \in \mathbf{R}_+$, $\tilde{Q}_t \ll P$ on $\mathcal{F}^0_{t+}$, or, equivalently, $\tilde{Q}_0 \ll P$ on $\mathcal{F}^0_{0+}$.

(9) **Theorem.** *A finite measure* Q *on* $\mathcal{F}^0_{0-}$ *which is progressively absolutely continuous may be written uniquely as* $Q = P^-_\alpha + \mu$ *where* P^-_α *is the restriction to* $\mathcal{F}^0_{0-}$ *of the Palm measure of an (integrable) predictable AF* α, *and the measure* μ *is concentrated on a polar set.*

The meaning of "polar" is this. Let ξ be a random variable on Ω; the random set $S_\xi(\omega) = \{t\colon \xi \circ \theta_t(\omega) \neq 0\}$ is called the *spoor* of ξ. If $\xi = I_A$, we speak of the "spoor of A". Then ξ (or A if $\xi = I_A$) is called polar (*thin*) if its spoor is a.s. empty (a.s. locally finite). Thus ξ is *polar* iff $\xi \circ \theta = (\xi \circ \theta_t(\omega))$ is an "evanescent" process [2]. Further, A is *semipolar* if it is contained in a countable union of thin sets. As indicated earlier, P^-_α further decomposes into a measure which charges no semipolar and a measure which

lives on a semipolar, but charges no polar, set (see §2). We note that if Q charges no polar set in $\mathcal{F}^0_{0-}$, then (using a Fubini argument) it is progressively absolutely continuous.

As an immediate consequence of (9) (using the trivial filtration $\mathcal{F}^0_t \equiv \mathcal{F}^0$ for (b)) we have

(10) **Theorem.** (a) *A finite measure Q on $\mathcal{F}^0_{0-}$ is the restriction to $\mathcal{F}^0_{0-}$ of the Palm measure of an integrable, predictable AF iff Q charges no polar set in $\mathcal{F}^0_{0-}$.*

(b) *A finite measure Q on $\mathcal{F}^0$ is a Palm measure iff Q charges no polar set in $\mathcal{F}^0$, in which case the corresponding AF (α_t) is the right-continuous regularization of (dQ^t/dP) where $Q^t(A) = \int_0^t Q(\theta_s A)\,ds$, $A \in \mathcal{F}^0$.*

Proof of (9). The uniqueness is clear, since (3) implies that a Palm measure charges no polar set. Let $Z_t \in (\mathcal{F}^0_{t+})_+$ be the Radon-Nikodym derivative $d\tilde{Q}_t/dP$ on $\mathcal{F}^0_{t+}$. Since $t \to EZ_t$ is right-continuous one may choose an a.s. right-continuous version $Z = (Z_t)$ and it is easy to see that Z is a potential, though not necessarily of class (D) (see [13] for terminology).

Now a change of variables yields $\tilde{Q}_t(A) = e^{-t}\tilde{Q}_0(\theta_t A)$, $A \in \mathcal{F}^0_{t+}$, from which follows immediately that, for each $t \in \mathbf{R}_+$, $Z_t = e^{-t}Z_0 \circ \theta_t$ a.s. By the remark following (5), we may assume Z_t is homogeneous: $Z_t \equiv e^{-t}Z_0 \circ \theta_t$.

It is well known that any potential Z has a unique decomposition $Z = N + Y$, where N is a local martingale (and also a potential) and Y is a potential of class (D). We show now that N and Y may be chosen homogeneous. Notice that $Z_{t+s}(\omega) = e^{-t}Z_s \circ \theta_t(\omega)$ for all ω, s, t. For t fixed, consider the following two decompositions of Z_{t+s}, $s \geq 0$.

$$Z_{t+s} = \begin{cases} N_{t+s} + Y_{t+s}, \\ e^{-t}N_s \circ \theta_t + e^{-t}Y_s \circ \theta_t. \end{cases} \tag{11}$$

It is tedious, but straightforward, to check that both N_{t+s} and $e^{-t}N_s \circ \theta_t$, $s \geq 0$, are local martingales and that both Y_{t+s} and $e^{-t}Y_s \circ \theta_t$, $s \geq 0$, are class (D) potentials, all relative to the σ-fields $\mathcal{F}_{t+s}$, $s \geq 0$. By uniqueness, the two expressions "match" correctly, and we conclude $N_{t+s} = e^{-t}N_s \circ \theta_t$ for all s, a.s., and similarly for Y_{t+s}. Putting $s = 0$, we find N, Y are "almost homogeneous" and may be replaced by homogeneous modifications, which we again denote by N and Y.

By the Doob-Meyer decomposition theorem [13, p. 119] we may write $Y = M - A$, where $M = (M_t)$ is a uniformly integrable martingale, and $A = (A_t)$

is a predictable (= natural) integrable increasing process. Since Y is homogeneous, we have two decompositions of Y_{t+s} for $s \geq 0$ (t fixed) analogous to (11):

$$Y_{t+s} = \begin{cases} M_{t+s} - A_{t+s} = M_{t+s} - A_t - (A_{t+s} - A_t), \\ e^{-t}M_s \circ \theta_t - e^{-t}A_s \circ \theta_t. \end{cases}$$

Noting that $e^{-t}M_s \circ \theta_t$ and $M_{t+s} - A_t$ are uniformly integrable martingales, and $e^{-t}A_s \circ \theta_t$, $A_{t+s} - A_t$ are predictable increasing processes (all relative to $\{\mathcal{F}_{t+s}\}$, $s \geq 0$, t fixed), we have by the uniqueness of the decomposition,

$$(12) \qquad A_{t+s} - A_t = e^{-t}A_s \circ \theta_t.$$

Now let $\alpha_t = \int_0^t e^s \, dA_s$. (By $\int_a^b$ we will always mean $\int_{(a,b]}$.) From (12) it follows easily that α is a predictable AF. (This argument was inspired by Maisonneuve [10].)

Now for any Palm measure, say P_β, we have (see [5])

$$(13) \; E\int_{\mathbf{R}} u(s, \omega)\beta(ds, \omega) = \int_\Omega \int_{\mathbf{R}} u(s, \theta_{-s}\omega) \, ds P_\beta(d\omega), \qquad u \in (\mathcal{B} \otimes \mathcal{F}^0)_+.$$

Hence for any $A \in \mathcal{F}^0_{t+}$,

$$E(Y_t; A) = E\left(\int_t^\infty e^{-s} \, d\alpha_s; A\right) = \int_t^\infty e^{-s} P_\alpha(\theta_s A) \, ds.$$

So we may write

$$(14) \; \int_t^\infty e^{-s} Q \circ \theta_s(A) \, ds = E(N_t; A) + \int_t^\infty e^{-s} P_\alpha \circ \theta_s(A) \, ds, \qquad A \in \mathcal{F}^0_{t+}.$$

Now define a measure $\tilde{\mu}$ on $\mathcal{P}^0$ by equation (8) with Q replaced by $\mu = Q - P_\alpha^-$, P_α^- being the restriction of P_α to $\mathcal{F}^0_{0-}$. Using (14), we see that $\tilde{\mu}$ is positive on sets of the form $(t, \infty) \times A$, $A \in \mathcal{F}^0_{t+}$, and hence on all of $\mathcal{P}^0$; moreover

$$(15) \qquad E(N_t; A) = \tilde{\mu}[(t, \infty) \times A] = \int_t^\infty e^{-s} \mu \circ \theta_s(A) \, ds, \qquad A \in \mathcal{F}^0_{t+}.$$

(The existence of a measure $\tilde{\mu}$ on $\mathbf{R}_+ \times \Omega$ satisfying the first equality in (15) is established by Föllmer [3] in a different situation.) It is easy to see that $\mu \geq 0$ so it only reamins to prove that μ lives on a polar set.

Having chosen N homogeneous, i.e. $N_t \equiv e^{-t}\overline{n} \circ \theta_t$ for an excessive function $\overline{n} \in (\mathcal{F}^0_{0+})_+$, note first that we may extend N_t to all $t \in \mathbf{R}$ and still have a supermartingale. Moreover, N_t will be right-continuous with finite left limits for all $t \in \mathbf{R}$ a.s. since $t \to \overline{n} \circ \theta_t$ has these properties.

Define, for each $n \geq 1$, $R_n = \inf\{r > 0, \text{ rational}: N_r > n\}$. Each $R_n \in \mathcal{S}^0_+$, the family of stopping times of $\{\mathcal{F}^0_{t+}\}$, $t \in \mathbf{R}_+$, and on right-continuous paths coincides with $\inf\{r > 0: N_r > n\}$. Starting with *discrete* $T \in \mathcal{S}^0_+$, one easily shows that the stochastic interval $]]T, \infty[[\in \mathcal{P}^0$ for *any* $T \in \mathcal{S}^0_+$, and hence $K = \bigcap_n]]R_n, \infty[[\in \mathcal{P}^0$. The argument in [3] shows that $\tilde{\mu}$ is supported by K and that K is evanescent. (The set K in [3] is slightly different, but, since $\tilde{\mu}$ puts no mass on $\{\infty\} \times \Omega$ because N_t is a potential, the argument given there goes through.)

Define $\xi(\omega) = \int_0^\infty e^{-r} 1_K(r, \theta_{-r}\omega)\, dr$; since $K \in \mathcal{P}^0$, (6) implies $\xi \in (\mathcal{F}^0_{0-})$. Clearly $\mu(\xi) = \tilde{\mu}(K)$, and $\mu(1 - \xi) = \tilde{\mu}(K^c) = 0$. Hence $\xi = 1$ μ-a.e. and $\mu\{\xi = 0\} = 0$, i.e. μ lives on the set $\{\xi > 0\}$. To show this set is polar, let G be the set of $\omega \in \Omega$ on which the trajectory $N_t(\omega)$ has finite left limits and is right-continuous at all $t \in \mathbf{R}$. Then G, and so G^c, is invariant, and $P(G^c) = 0$. Suppose $\xi(\omega) > 0$. Then $(r, \theta_{-r}\omega) \in K$ for some $r > 0$, i.e. $R_n(\theta_{-r}\omega) < r$ for all $n \geq 1$, and this puts ω in G^c. But $\{\xi > 0\} \subset G^c$ implies $\{\xi > 0\}$ is polar since G^c is an invariant null set. This completes the proof of (9).

We now sketch the proof of another decomposition for an arbitrary (finite) Q on $\mathcal{F}^0_{0-}$, which is valid under the additional assumption that $\{\mathcal{F}^0_t\}$ is a *standard system* [3], [15], which means

(a) each $\mathcal{F}^0_t$ is σ-isomorphic to the Borel σ-field of a Polish space;

(b) for any increasing sequence t_n, and decreasing sequence of sets A_n, such that A_n is an atom of $\mathcal{F}^0_{t_n}$, we have $\bigcap_n A_n \neq \emptyset$.

Unfortunately, the usual filtrations on the standard spaces of flow theory, such as $\mathfrak{C}$ and $\mathfrak{W}$ previously mentioned, are not standard in the above sense. We will indicate later how to circumvent this difficulty for those two cases.

Let Q be a finite measure on $\mathcal{F}^0_{0-}$ and define $\tilde{Q}$ as before (see (8)). For each $t \in \mathbf{R}_+$, the measure $\tilde{Q}_t$ has a Lebesgue decomposition on $\mathcal{F}^0_{t+}$, namely $\tilde{Q}_t(A) = Q'_t(A) + Q''_t(A)$, with $Q'_t \ll P$ on $\mathcal{F}^0_{t+}$, and $Q''_t \perp P$ on $\mathcal{F}^0_{t+}$ ($\perp$ means "singular"). An easy argument using the uniqueness of the Lebesgue decomposition shows that $Q'_t = e^{-t} Q'_0 \circ \theta_t$ and $Q''_t = e^{-t} Q''_0 \circ \theta_t$. Let $Z_t = dQ'_t/dP$ on $\mathcal{F}^0_{t+}$. We may choose a homogeneous version of the potential $Z = (Z_t)$ just as in the proof of (9), and this splits into a local martingale N plus a class (D) potential Y, both of which are homogeneous. Now let $\tilde{\mu}$ be the *Föllmer measure* of the local martingale N, i.e. the unique measure on $\mathcal{P}^0$ such that the first equality in (15) holds. We thus have, α being as in the proof of (9),

$$\tilde{Q} = \tilde{Q}_\alpha + \tilde{\mu} + \tilde{\nu} \tag{16}$$

where $\tilde{Q}_a$ is defined by the right-hand side of (8) with P_a in place of Q, $\tilde{\nu} \equiv \tilde{Q} - \tilde{Q}_a - \tilde{\mu}$. In this way, $\tilde{\nu}[(t, \infty) \times A] = Q_t''(A)$, $A \in \mathcal{F}^0_{t+}$.

Equation (16) exhibits $\tilde{Q}$ as the sum of the progressively absolutely continuous measure $\tilde{M} = \tilde{Q}_a + \tilde{\mu}$ and the "progressively singular" measure $\tilde{\nu}$. One establishes easily that such a decomposition is unique.

Define measures μ, ν on $\mathcal{F}^0_{0-}$ by $\mu(A) = \tilde{\mu}(I_A \circ \theta)$, $\nu(A) = \tilde{\nu}(I_A \circ \theta)$.

(17) **Lemma.** *For every* $u \in (\mathcal{P}^0)_+$,

(18) $$\tilde{\mu}(u) = \int_\Omega \int_0^\infty e^{-s} \theta^- u(s, \omega)\, ds\, \mu(d\omega)$$

and similarly for $\tilde{\nu}$ *and* ν.

Proof. Define a transformation T_t on $\mathcal{B}_+ \otimes \mathcal{F}^0$ for each $t \in \mathbf{R}_+$ by $T_t u(s, \omega) = u(s + t, \theta_{-t}\omega)$. A monotone class argument shows that T_t: $(\mathcal{P}^0)_+ \to (\mathcal{P}^0)_+$. From (8) we find that

(19) $$\tilde{Q}(T_t u) = e^t \tilde{Q}(I_{[[t,\infty[[} u), \qquad t \in \mathbf{R}_+, \quad u \in (\mathcal{P}^0)_+.$$

By looking at the generators of $\mathcal{P}^0$, we also find that $\tilde{M}(T_t u)$ defines a progressively absolutely continuous measure, while $\tilde{\nu}(T_t u)$ is progressively singular (t fixed). The uniqueness of the decomposition (16) shows that (19) holds for $\tilde{M}$ (resp. $\tilde{\nu}$) in place of $\tilde{Q}$; since $\tilde{Q}_a$ also satisfies (19), so does $\tilde{\mu}$.

In view of (6), it will suffice to verify (18) for $u = \theta^+(I_{[t,\infty)}\xi)$, $t \geq 0$, $\xi \in (\mathcal{F}^0_{0-})_+$, in which case the right-hand side of (18) reduces to $e^{-t}\tilde{\mu}(\xi \circ \theta)$. The left-hand side is

$$\begin{aligned} \tilde{\mu}(I_{[[t,\infty[[}\theta^+\xi) &= e^{-t}\tilde{\mu}(T_t(\xi \circ \theta)) \quad \text{(by (19))} \\ &= e^{-t}\tilde{\mu}(\xi \circ \theta). \end{aligned}$$

The fact that μ is carried by a polar set is proven just as before and we can state, *given that* $\{\mathcal{F}^0_t\}$ *is standard*:

(20) **Theorem.** *A finite measure* Q *on* $\mathcal{F}^0_{0-}$ *may be written as* $Q = P^-_a + \mu + \nu$, *where* P^-_a *and* μ *are as described in* (9), *and* ν *is such that*

$$\tilde{\nu}(u) = \int_\Omega \int_0^\infty e^{-s}\theta^- u(s, \omega)\, ds\, \nu(d\omega), \qquad u \in (\mathcal{P}^0)_+,$$

is progressively singular.

As we indicated, the spaces $\mathfrak{C}$, $\mathfrak{D}$ and $\mathfrak{W}$ are *not* standard. To illustrate how to overcome this difficulty, introduce the space $\mathfrak{C}'$ consisting of all functions f: $\mathbf{R} \to \mathbf{R} \cup \{\Delta\}$, where $\Delta \notin \mathbf{R}$ is an adjoined "death point" such that f is continuous on $\mathbf{R}$ for all $t < \zeta$ and $f(t) = \Delta$ for all $t \geq \zeta$. The "lifetime" $\zeta \leq \infty$ depends on f. Clearly $\mathfrak{C} \subset \mathfrak{C}'$, and $f \in \mathfrak{C}$ iff $\zeta = +\infty$. Again let

θ_t "shift" f by t, $X_t f = f(t)$, and define $\mathcal{F}'_t = \sigma\{X_s, s \leq t\}$, all relative to $\mathfrak{C}'$. In this way, $\mathcal{F}^0_t = \mathcal{F}'_t \cap \mathfrak{C}$, $\mathfrak{C}$ is an invariant subset of $\mathfrak{C}'$, and a stationary measure P on $\mathfrak{C}$ extends naturally to $\mathfrak{C}'$ by $P'(A) = P(A \cap \mathfrak{C})$, $A \in \mathcal{F}' = \bigvee_t \mathcal{F}'_t$. The filtration $\{\mathcal{F}'_t\}$ is standard. (To "standardize" $\mathfrak{B}$, one introduces $\mathfrak{B}'$ consisting of all nonempty subsets $w \in \mathbf{R}$ which are locally finite before $\zeta = \sup w \leq \infty$. We only pursue the case $\mathfrak{C}'$; the argument for $\mathfrak{B}$ is entirely similar.)

Let Q be a finite measure on $\mathcal{F}^0_{0-} = \mathfrak{C} \cap \mathcal{F}'_{0-}$, and extend it to $\mathcal{F}'_{0-}$ by $Q'(A) = Q(A'\mathfrak{C})$, $A \in \mathcal{F}'_{0-}$. By (20)

$$Q' = P^-_\alpha + \mu + \nu, \tag{21}$$

all measures on $\mathcal{F}'_{0-}$. Restricting each of these to $\mathcal{F}^0_{0-} \subset \mathcal{F}'_{0-}$ we obtain a decomposition of Q; it only remains to show it is of the desired form. Now P^-_α kills every polar set in $\mathcal{F}^0_{0-}$ since every polar set in $\mathcal{F}^0$ relative to P is also polar relative to P'; indeed, the restriction to $\mathfrak{C}$ of P_α is the Palm measure of the restriction of α to $\mathfrak{C}$. As for the restriction to $\mathcal{F}^0_{0-}$ of μ, it is obvious that it lives on $N\mathfrak{C}$, where $N \in \mathcal{F}'_{0-}$ is polar and $\mu(N^c) = 0$, and that $N\mathfrak{C} \in \mathcal{F}^0_{0-}$ is polar. One also can show that ν restricted to $\mathcal{F}^0_{0-}$ is as in (20).

Finally we remark that, if we write the Lebesgue and $\tilde{Q}'_t(A')$ (with the obvious notation),

$$\tilde{Q}_t(A) = E(Z_t; A) + K_t(A), \qquad A \in \mathcal{F}^0_{t+},$$
$$\tilde{Q}'_t(A') = E'(Z'_t; A') + K'_t(A'), \qquad A' \in \mathcal{F}'_{t+},$$

where K_t, K'_t are P-singular (resp. P'-singular), and Z_t, Z'_t are the Radon-Nikodym derivatives of the P- (P'-) absloutely continuous pieces, then Z, Z' are potentials and the pieces match properly. Indeed, $K'_t(A') = K_t(A'\mathfrak{C})$, and Z'_t may be taken as the "canonical extension" of Z_t to $\mathfrak{C}'$: first choose the homogeneous version of Z, and then let $Z'_t(f) = 0$ if $t \geq \zeta$ and $Z'_t(f) = Z_t(\tilde{f})$ if $t < \zeta$, where $\tilde{f} \in \mathfrak{C}$ is any function agreeing with $f \in \mathfrak{C}'$ for all $s \leq t + \epsilon$ for sufficiently small $\epsilon > 0$. The $\mathcal{F}^0_t$-measurability of Z_t guarantees the definition to be independent of the choice of $\tilde{f}$, and Z' is also homogeneous.

It is an open question whether every filtered dynamical system can be embedded in a standard system as in the above examples. We should also mention that the measure ν in (20) is something of a mystery to us.

We conclude this section by pointing out two applications of our results. Let $A = (A_t)$, $t \in \mathbf{R}_+$, be an integrable, increasing process relative to $\{\mathcal{F}_t\}$, that is (see [2]), $A_0 \equiv 0$, A is right-continuous, nondecreasing and $EA_\infty < \infty$.

Define a measure μ_A on $R_+ \times \Omega$ by

$$\mu_A(u) = E\int_0^\infty u(s, \omega)\, dA_s(\omega), \qquad u \in (\mathcal{B}_+ \otimes \mathcal{F}^0)_+,$$

and a (finite) measure Q on $\mathcal{F}^0$ by $Q(\xi) = \mu_A(\xi \circ \theta)$. Obviously, Q kills every polar set in $\mathcal{F}^0$, and hence by (10) is the Palm measure of an (integrable) AF α. One easily checks that, in fact,

$$\alpha_t(\omega) = \int_R \int_0^\infty I_{(0,t]}(s + r)\, dA_r(\theta_s \omega)\, ds, \qquad t \in R_+.$$

We call α the *additive projection* of A. A special case of this is used in [6], and Mecke [12] has noted a similar idea.

The following result has a Markovian analogue [16].

(22) **Theorem.** *A set $N \in \mathcal{F}^0$ is polar if and only if it is charged by no Palm measure.*

Proof. If N is polar, we have already observed that $P_\alpha(N) = 0$ for every AF α. Suppose N is not polar. Its spoor $S_N(\omega)$ is then nonempty on a set of ω having positive probability; in fact, for almost every $\omega \in \Omega$ for which $S_N(\omega) \neq \emptyset$, we shall see that $S_N(\omega)$ is unbounded above. Observe, to begin with, that $S_N(\omega)$ is a "homogeneous set" in the sense that for every $t \in R$, $S_N(\theta_t \omega) = S_N(\omega) - t$. Let $B_N = \{\sup S_N > n\}$. Since the projection on Ω of sets in $\mathcal{B} \otimes \mathcal{F}$ is in $\mathcal{F}$ (see [2, I., T32]), $B_n \in \mathcal{F}$ for each $n \geq 1$. The ergodic theorem now yields

$$\lim_{t\to\infty} t^{-1}\int_0^t I_{B_n}(\theta_{-s}\omega)\, ds = P(B_n \mid \mathcal{A}) \quad \text{a.s.}$$

where $\mathcal{A}$ denotes the θ-invariant σ-field in Ω. Suppose $S_N(\omega) \neq \emptyset$. Then $S_N(\theta_{-s}\omega) = S_N(\omega) + s$ has supremum greater than n when s is sufficiently large. Hence if $B = \{S_N \neq \emptyset\}$, we have $P(B_n \mid \mathcal{A}) = 1$ a.s. on B. Since B is invariant, $P(B \cap B_n^c) = 0$, which proves our point.

Now set $D(\omega) = S_N(\omega) \cap (0, \infty) \in \mathcal{B}_+$; clearly $P(D \neq \emptyset) = P(B) > 0$. According to [2, I., T37], there exists a nonnegative random variable $\tau \in (\mathcal{F})_+$ such that $\tau(\omega) \in D(\omega)$ for $D(\omega) \neq \emptyset$ and otherwise $\tau = \infty$. Define an integrable, increasing process $A = (A_t)$ by $A_t(\omega) = I_{\{\tau \leq t\}}(\omega)$; A is flat, except for a unit jump at τ if $\tau < \infty$. We then have

$$\mu_A(I_N \circ \theta) = E\int_0^\infty I_N \circ \theta_s(\omega)\, dA_s(\omega) = E(I_N \circ \theta_\tau;\, D \neq \emptyset) = P(B) > 0.$$

By our work above, $P_\alpha(N) = \mu_A(I_N \circ \theta) > 0$, where α is the additive projection of A, and (22) is proven.

Note. A similar argument, using the material in [2, Chapter VI (esp. §3 and T37)], shows that *a set $N \in \mathcal{F}^0$ has an a.s. countable spoor $S_N(\omega)$ iff $P_\alpha(N) = 0$ for every continuous AF α.*

For our second application, let $X = (X_t)$, $t \in \mathbf{R}$, be a strictly stationary measurable process such that $X_t = X_0 \circ \theta_t$, and assume X is adapted to a filtration $\{\mathcal{F}_t^0\}$. Denote by $(E, \mathcal{E})$ the state space of X, with $\mathcal{E}$ assumed separable, and let $\pi(\Gamma) = P\{X_t \in \Gamma\}$ (independent of t) be the one-dimensional distribution of the process. We say that X has a *local time* if there exist AF's $\alpha^x = (\alpha_t^x)$, $x \in E$, such that, for almost every $\omega \in \Omega$,

$$(23) \qquad \int_\Gamma \alpha_t^x(\omega)\pi(dx) = \int_0^t I_\Gamma(X_s(\omega))\, ds \quad \text{for all } t \in \mathbf{R}_+,\ \Gamma \in \mathcal{E}.$$

(Suitable measurability restrictions must be imposed; we omit the details.) Let $\{P^x\}$, $x \in E$, be a regular conditional probability given X_0, i.e. a family of measures on $\mathcal{F}^0$ such that, for every $A \in \mathcal{F}^0$, $\Gamma \in \mathcal{E}$, $P(A, X_0 \in \Gamma) = \int_\Gamma P^x(A)\pi(dx)$. We know [4] a local time exists iff P^x is a Palm measure for π-a.e. $x \in E$ (in which case the AF's are predictable for a.e. x).

Suppose only that there exist AF's α^x which are *predictable* and such that $P^x = P_{\alpha^x}$ on $\mathcal{F}_{0-}^0$ for a.e. x. Set

$$\beta_t^\Gamma(\omega) = \int_0^t I_\Gamma(X_s(\omega))\, ds, \quad \alpha_t^\Gamma(\omega) = \int_\Gamma \alpha_t^x(\omega)\pi(dx),$$

for $t \in \mathbf{R}_+$, $\Gamma \in \mathcal{E}$. An easy computation shows that β^Γ and α^Γ have the same Palm measure on $\mathcal{F}_{0-}^0$, hence on all of $\mathcal{F}^0$ since β^Γ and α^Γ are both predictable. Consequently $\alpha_t^\Gamma(\omega) = \beta_t^\Gamma(\omega)$ for all t, a.s., which yields (using separability of $\mathcal{E}$) $\alpha_t^\Gamma(\omega) = \beta_t^\Gamma(\omega)$ for all t and all Γ, a.s. Thus we have proven

(24) **Theorem.** *A local time exists iff, for almost every $x \in E$, the measure P^x charges no polar set in $\mathcal{F}_{0-}^0$.*

Notice, however, that "polar" is defined in terms of sets in Ω which are avoided by the *flow* θ_t rather than those in E which are avoided by the *process* X_t.

We conclude with an example, based on a construction due to Maisonneuve [10], of the "local time" of an arbitrary random set. Suppose, for each $\omega \in \Omega$, we are given a Borel set $M(\omega)$ of $\mathbf{R}$, homogeneous in that $M(\theta_t\omega) = M(\omega) - t$ for all $t \in \mathbf{R}$, $\omega \in \Omega$. (For example, $M(\omega) = \{t : X_t(\omega) = 0\}$ with (X_t) as above.) Define $\tau = \inf(M \cap (0, \infty))$ (or $\tau = \infty$ if $M \cap (0, \infty)$ is empty), and $\tau_t(\omega) = t + \tau \circ \theta_t(\omega)$. The random variable τ is "terminal": $\tau = \tau_t$ whenever $\tau > t$.

Let $Z = E(e^{-\tau} | \mathcal{F}_0)$. Then Z is excessive, and by (5) we may choose a nice version of Z so that $Z_t = e^{-t} Z \circ \theta_t$ is a homogeneous potential, obviously of class (D). Proceeding as in the proof of (9) we write the decomposition $Z_t = M_t - A_t$, and let $\alpha_t = \int_0^t e^s\, dA_s$. Then α is a predictable AF. Under further conditions on M, $\alpha(dt, \omega)$ is a.s. carried by $M(\omega)$. The Palm measure of α is

$$E_\alpha(\xi) = -E\int_0^\infty \xi^* \circ \theta_s \, de^{-\tau_s}, \qquad \xi \in (\mathcal{F}^0)_+,$$

where ξ^* is the "predictable projection" of ξ. Put another way, α is just the predictable projection of $-e^t de^{-\tau_t}$ (see §2). If $X = (X_t)$ has a local time, say α^x, we do not know the connection, if any, between α^x and the local time of $M^x(\omega) = \{t\colon X_t(\omega) = x\}$, $x \in E$, at least outside the Markov case.

2. **Characterization of additive functionals.** With assumptions (I), (II) of §1 still in force, we will need the following basic fact, borrowed from [8] (see also [9]).

(25) **Theorem.** *For every $\xi \in (\mathcal{F}^0)$ which is either bounded or nonnegative, there exists $\xi^\# \in (\mathcal{F}^0_{0+})$ (resp. $\xi^* \in (\mathcal{F}^0_{0-})$) such that $(\xi^\# \circ \theta_t)$, $t \in \mathbf{R}$ (resp. $(\xi^* \circ \theta_t)$) is the well-measurable (resp. predictable) projection of the process $(\xi \circ \theta_t)$, $t \in \mathbf{R}$.*

Moreover, $\xi^\#$ (resp. ξ^*) is bounded or nonnegative with ξ and is unique up to a polar function. We note that the process $\xi^* \circ \theta$ is actually in $(\mathcal{P}^0)$ by (7)(b) while the notions of well-measurability, etc. refer to the family $\{\mathcal{F}_t\}$.

Our results in this section will be of two kinds: the first type classifies an AF α in accordance with the behavior of P_α under projection, while, in the second type, we give conditions under which, for example, the dual predictable projection of α is a.s. absolutely continuous. These latter results are generalizations of some of the work of Papangelou [14].

Before going on, we recall some material from the general theory of processes [2]. Let $u = (u(t, \omega))$ be a process and $A = (A_t(\omega))$ an increasing process, $EA_t < \infty$, $t \in \mathbf{R}_+$, and $\{\mathcal{F}_t\}$ an increasing family of σ-fields on Ω which is right-continuous and with each $\mathcal{F}_t$ completed by all P-null sets. We write $\mathcal{W}$, $\mathcal{P}$ for the well-measurable (resp. predictable) σ-fields on $\mathbf{R}_+ \times \Omega$, and note that $\mathcal{P} \subset \mathcal{W}$. The accessible σ-field falls between $\mathcal{P}$ and $\mathcal{W}$, but will be omitted from our discussion.

Writing $w(u)$ and $p(u)$ for the well-measurable and predictable projections of the process u, the dual well-measurable (resp. predictable) projection of the increasing process A is defined as follows: A^w (resp. A^p) is the unique well-measurable (resp. predictable) increasing process such that

$$(26) \qquad E\int_0^\infty w(u)(s, \omega)\, dA_s(\omega) = E\int_0^\infty u(s, \omega)\, dA^w_s(\omega), \qquad u \in (\mathcal{B}_+ \otimes \mathcal{F})_+,$$

and similarly for A^p. For an increasing process, we note that well-measurability is equivalent to being adapted.

For an integrable RAF α, we now denote by $\alpha^\#$ (resp. α^*) the dual well-measurable (resp. predictable) projection as defined above.

(27) **Theorem.** *The increasing processes* $\alpha^{\#}$, α^* *are AF's whose Palm measures are*

$$E_{\alpha^{\#}}(\xi) = E_\alpha(\xi^{\#}), \quad E_{\alpha^*}(\xi) = E_\alpha(\xi^*), \qquad \xi \in (\mathcal{F}^0)_+. \tag{28}$$

Proof. It suffices to treat the predictable case, the other being entirely analogous, even somewhat easier. Suppose, for the moment, that α^* is an AF. Let $\xi \in (\mathcal{F}^0)_+$. Then $p(\xi \circ \theta) = \xi^* \circ \theta$, and the predictable version of (26) gives

$$E\int_0^\infty e^{-s}\xi \circ \theta_s \alpha^*(ds) = E\int_0^\infty e^{-s}\xi^* \circ \theta_s \alpha(ds), .$$

i.e. (28) holds. To show α^* is an AF, it is enough to establish

$$E[\alpha^*_{t+s} - \alpha^*_t;\ A] = E[\alpha^*_s \circ \theta_t;\ A], \qquad A \in \mathcal{F}. \tag{29}$$

The left side of (29) can be written as

$$\begin{aligned} E\int_0^\infty I_{(t,t+s]}(r) I_A(\omega)\alpha^*(dr,\omega) &= E\int_0^\infty p(I_{(t,t+s]} I_A)\, d\alpha \\ &= E\int_0^\infty I_{(t,t+s]}(r) P(A\,|\,\mathcal{F}_{r-})\alpha(dr) \\ &= E\int_0^s P(A\,|\,\mathcal{F}_{(r+t)-})\alpha(dr, \theta_t\omega), \end{aligned}$$

where $P(A|\mathcal{F}_{r-})$ denotes the left-continuous modification of the martingale $P(A|\mathcal{F}_r)$. Using stationarity of the flow, it is easy to prove that, for each r, t, $P(A|\mathcal{F}_{(t+r)-}) = P(\theta_t A|\mathcal{F}_{r-}) \circ \theta_t$ a.s., and so, fomally, the last displayed expression becomes

$$= E\int_0^s P(\theta_t A|\mathcal{F}_{r-})\alpha(dr) = E\int_0^\infty I_{\theta_t A}\alpha^*(dr) = E[\alpha^*_s \circ \theta_t;\ A].$$

The problem is to show that $P(\theta_t A|\mathcal{F}_{r-}) \circ \theta_t$ may be chosen *indistinguishable* from $P(A|\mathcal{F}_{(t+r)-})$, as r varies. However, both processes are a.s. left-continuous in r, hence are indistinguishable. Q.E.D.

Since Palm measures determine AF's (up to indistinguishability):

(30) **Corollary.** *An AF* α *is adapted (resp. predictable) iff* $E_\alpha(\xi) = E_\alpha(\xi^{\#})$ *(resp.* $E_\alpha(\xi) = E_\alpha(\xi^*)$*) for every* $\xi \in (\mathcal{F}^0)_+$.

Notes. (1) If α is adapted, then α will be predictable iff $E_\alpha(\xi) = E_\alpha(\xi^*)$ for all $\xi \in (\mathcal{F}^0_{0+})_+$ since the Palm measure of an adapted AF is completely determined by its action on $\mathcal{F}^0_{0+}$.

(2) Since P^-_α kills polar sets in $\mathcal{F}^0_{0-}$, there exists (by (10)(a)) a *predictable* AF β such that $P^-_\beta = P^-_\alpha$; in fact, $\beta = \alpha^*$ since $E_\beta(\xi) = E_\beta(\xi^*) = E_\alpha(\xi^*) = E_{\alpha^*}(\xi)$ for $\xi \in (\mathcal{F}^0)_+$.

(31) **Corollary.** *Additive projection (see §1) preserves well-measurability (resp. predictability).*

We consider next the splitting of an AF α into the sum of a continuous AF α_c and a purely discrete AF α_d, i.e. the measure $\alpha_c(dt, \omega)$ has no atoms for each $\omega \in \Omega$, whereas $\alpha_d(dt, \omega)$ is the sum of countably many point masses depending on $\omega \in \Omega$. The corresponding Palm measures are denoted P_c, P_d. Given α, α_c and α_d are obtained from the usual decomposition of a measure into a continuous plus a discrete piece. If α is adapted, α_c and α_d will be likewise, and α_c will be predictable; if α is predictable, α_d will be also. Let $A \in \mathcal{F}^0$ have an a.s. countable spoor S_A. Clearly $P_c(A) = 0$. In particular, this is the case when A is semipolar.

Now consider the discrete part α_d. Denote the mass on $\{0\}$ by Δ: $\Delta(\omega) = \alpha_d(0, \omega) - \alpha_d(0-, \omega) = -\alpha_d(0-, \omega)$ (there is no restriction (see §0) in assuming $\Delta \in (\mathcal{F}^0)_+$); also let $\Delta_t(\omega) = \Delta \circ \theta_t(\omega)$ be the mass on $\{t\}$. It is well known that P_d is supported on the set $\Omega_d = \{\Delta > 0\}$. Writing $\Omega_d = \bigcup_{n=1}^{\infty}\{\Delta > 1/n\}$ and recalling that α_d is finite on compacts, we see that Ω_d is semipolar.

When α is predictable, we get a finer decomposition: we can then show that P_d is supported by a semipolar set *in* $\mathcal{F}^0_{0-}$, and this in turn will lead to the decomposition promised in §1.

(32) **Lemma.** *The Palm measure of a purely discrete, predictable AF α is carried by a semipolar set in $\mathcal{F}^0_{0-}$.*

Proof. We choose for α the "perfect version" described in §0. The process $\Delta \circ \theta_t = \alpha(t) - \alpha(t-)$ is then predictable, and is therefore indistinguishable from its predictable projection $\Delta^* \circ \theta_t$. It follows that the set $N^* = \{\Delta^* > 0\}$ (which is in $\mathcal{F}^0_{0-}$) is semipolar since its spoor is a.s. the same as the spoor of Ω_d, and $P_\alpha(\Delta^* = 0) = 0$, since Palm measures fail to distinguish indistinguishable processes. The set N^* is thus the one required by the theorem.

The following is now immediate upon recalling our remarks in §1:

(33) **Theorem.** *A finite measure Q as in* (9) *has a decomposition $Q = P_c^- + P_d^- + \mu$ where P_c^- charges no semipolar set in $\mathcal{F}^0_{0-}$, P_d^- lives on a semipolar, but charges no polar set in $\mathcal{F}^0_{0-}$, and μ lives on a polar set in $\mathcal{F}^0_{0-}$.*

As a consequence of (33), we obtain the following refinements of the results in §1:

(34) **Corollary.** (a) *A finite measure Q on $\mathcal{F}^0$ is the Palm measure of*

a continuous AF iff Q charges no semipolar set.

(b) *For an AF α, α^* is continuous iff P_α charges no semipolar set in $\mathcal{F}^0_{0-}$.*

Part (a) follows by taking the trivial filtration $\mathcal{F}^0_t \equiv \mathcal{F}^0$ in (33), and part (b) from (33) and the note following (30).

A similar situation obtains for discrete AF's:

(35) **Corollary.** *An AF α is purely discrete iff its Palm measure is carried by a semipolar set.*

If N is semipolar, it is contained in the union of thin sets B_n, so we may consider N thin. By the argument in the proof of (22), the spoor $S_N(\omega)$ will be unbounded in both directions a.s. Since N is thin, we may enumerate the points of $S_N(\omega)$: $\ldots R_{-1}(\omega) < R_0(\omega) \leq 0 < R_1(\omega) < \ldots$, and we have $R_{n+1} = R_n + R_1 \circ \theta_{R_n}$ for each integer n. Define $\nu(dt, \omega)$ as the measure which puts unit mass on each $R_n(\omega)$. We leave it to the reader to check that ν is an AF and that $\alpha(dt, \omega) \ll \nu(dt, \omega)$ for almost every $\omega \in \Omega$ (to show this it suffices to check $Q \ll P_\nu$). The general semipolar case is an easy extension of this method.

We conclude with a characterization of the absolute continuity of α^* in terms of P_α similar to that given in [14] for point processes.

(36) **Theorem.** *The following three statements are equivalent:*

(a) *α^* is a.s. absolutely continuous (i.e. $\alpha^*(dt, \omega) \ll dt$);*

(b) *$P_\alpha \ll P$ on $\mathcal{F}^0_{0-}$;*

(c) *$t^{-1}E(\alpha(t)|\mathcal{F}_0)$ converges in L^1-norm as $t \downarrow 0$.*

Proof. Suppose $\alpha^*(dt, \omega) = \xi(t, \omega)\,dt$. Then, for any $A \in \mathcal{F}^0_{0-}$, $P_\alpha(A) = P_{\alpha^*}(A) = E[\int_0^1 \xi(t, \theta_{-t}\omega)\,dt;\, A]$ which proves (a) $\Rightarrow$ (b). Conversely, assuming (b), we can write $dP^-_\alpha = \xi\,dP$ for some $\xi \in (\mathcal{F}^0_{0-})_+$. Now for any $\eta \in (\mathcal{F}^0)_+$, $E_{\alpha^*}(\eta) = E_\alpha(\eta^*) = E(\xi\eta^*)$ since $\eta^* \in (\mathcal{F}^0_{0-})_+$. If $\eta = 0$ a.s., the same is true of η^*, hence $P_{\alpha^*} \ll P$ and we can conclude that $\alpha^*(dt, \omega) = \xi \circ \theta_t(\omega)\,dt$, since both are predictable and have the same Palm measure on $\mathcal{F}^0_{0-}$. Thus (b) $\Rightarrow$ (a).

We next show that (a) is equivalent to (c). Recall the *local ergodic theorem* [9] which states that, if $\xi \in L^1(\Omega, \mathcal{F}^0, P)$, $t^{-1}\int_0^t \xi \circ \theta_s\,ds \to \xi$ (as $t \downarrow 0$) a.s. and in L^1. Now from the definition of α^* we have $E(\alpha(t)|\mathcal{F}_0) = E(\alpha^*(t)|\mathcal{F}_0)$ (see [2, VT 37]), and, assuming (a), we have

$$t^{-1}E(\alpha(t)\,|\,\mathcal{F}_0) = t^{-1}E\left(\int_0^t \xi \circ \theta_s\,ds\,|\,\mathcal{F}_0\right) \tag{37}$$

for some $\xi \in L^1$. But $t^{-1}\int_0^t \xi \circ \theta_s\, ds \to \xi (L^1)$, so the right member of (37) converges in L^1 to ξ as well.

Conversely, assume that $t^{-1}E(\alpha(t)|\mathfrak{F}_0)$ converges in L^1 to some ξ, which we may take in $\mathfrak{F}^0_{0-}$. Again using $E[\alpha(t) - \alpha(s)|\mathfrak{F}_s] = E[\alpha^*(t) - \alpha^*(s)|\mathfrak{F}_s]$, $s \le t$, we will show

$$E\int_0^1 Y_s \alpha^*(ds) = E\int_0^1 Y_s \xi \circ \theta_s\, ds \tag{38}$$

for every continuous, adapted, bounded process Y_s. Equation (38) then extends to all predictable processes Y; since $\alpha^*(t)$ and $\int_0^t \xi \circ \theta_s\, ds$ are each predictable, (38) implies $\alpha^*(t) = \int_0^t \xi \circ \theta_s\, ds$ for all t, a.s.

Before proving (38), we require

(39) **Lemma.** *For* $0 \le \xi \in L^1$ *and* Y *as described in* (38),

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} Y_{k/n}\, \xi \circ \theta_{k/n} = \int_0^1 Y_s \xi \circ \theta_s\, ds \quad (in\ L^1). \tag{40}$$

The L^1-norm of the difference of the two members of (40) may be written

$$\left\| n^{-1} \sum_0^{n-1} \left(Y_{k/n} \xi \circ \theta_{k/n} - n\int_{k/n}^{(k+1)/n} Y_s \xi \circ \theta_s\, ds \right) \right\| \le C_n + D_n,$$

where

$$C_n = \left\| n^{-1} \sum_0^{n-1} Y_{k/n} \left(\xi \circ \theta_{k/n} - n\int_{k/n}^{(k+1)/n} \xi \circ \theta_s\, ds \right) \right\|,$$

$$D_n = \left\| \sum_0^{n-1} \int_{k/n}^{(k+1)/n} (Y_{k/n} - Y_s) \xi \circ \theta_s\, ds \right\|.$$

Let $|Y_s(\omega)| \le M < \infty$ for all s, ω. Then

$$C_n \le \frac{M}{n} \sum_0^{n-1} \left\| \xi - n\int_0^{1/n} \xi \circ \theta_s\, ds \right\| \to 0$$

by the local ergodic theorem. Also,

$$D'_n \equiv \left| \sum_0^{n-1} \int_0^{1/n} (Y_{k/n} - Y_{k/n+s}) \xi \circ \theta_s \circ \theta_{k/n}\, ds \right|$$

$$\le 2M \sum_0^{n-1} \int_0^{1/n} \xi \circ \theta_{s+k/n}\, ds = 2M \int_0^1 \xi \circ \theta_s\, ds$$

since $\xi \ge 0$; hence D'_n is dominated by an L^1 function. Now, for $\omega \in \Omega$ fixed, $Y(\omega)$ is uniformly continuous on $[0, 1]$; hence, given $\epsilon > 0$, $|Y_{k/n}(\omega) -$

$Y_{k/n+s}(\omega)| \leq \epsilon$ for all sufficiently large n, all $k \leq n-1$, and all s in $[0, 1/n]$. Thus $D_n' \leq \epsilon \int_0^1 \xi \circ \theta_s \, ds$ for n large; by dominated convergence, $D_n \to 0$ and (39) is proven.

Returning to the proof of (38), we have only to show

$$E \int_0^1 Y_s \alpha^*(ds) = \lim_{n \to \infty} E n^{-1} \sum_0^{n-1} Y_{k/n} \xi \circ \theta_{k/n}. \tag{41}$$

Let

$$A_n' = \left| \int_0^1 Y_s \alpha^*(ds) - \sum_0^{n-1} Y_{k/n}(\alpha^*(k+1)/n - \alpha^*(k/n)) \right|$$

$$= \left| \sum_0^{n-1} \int_{k/n}^{(k+1)/n} (Y_s - Y_{k/n}) \alpha^*(ds) \right|.$$

Since $|Y| \leq M$, we find $A_n' \leq 2M\alpha^*(1) \in L^1$; on the other hand, given $\epsilon > 0$, $A_n' \leq \epsilon \alpha^*(1)$ for all sufficiently large n, by a uniform continuity argument such as the one above. Hence by dominated convergence $EA_n' \to 0$ and we conclude

$$\int_0^1 Y_s \alpha^*(ds) = \lim_{n \to \infty} \sum_0^{n-1} Y_{k/n}(\alpha^*(k+1)/n - \alpha^*(k/n)) \quad (\text{in } L^1). \tag{42}$$

Now $\alpha^*(k+1)/n - \alpha^*(k/n) = \alpha^*_{1/n} \circ \theta_{k/n}$. Also

$$\left| E\left[\sum_0^{n-1} Y_{k/n} \alpha^*_{1/n} \circ \theta_{k/n} - n^{-1} \sum_0^{n-1} Y_{k/n} \xi \circ \theta_{k/n} \right] \right|$$

$$= \left| E\left[\sum_0^{n-1} Y_{k/n} E(\alpha^*_{1/n} | \mathcal{F}_0) \circ \theta_{k/n} - n^{-1} \sum_0^{n-1} Y_{k/n} \xi \circ \theta_{k/n} \right] \right|$$

$$= \left| \sum_0^{n-1} E Y_{k/n} \circ \theta_{-k/n} (E(\alpha^*_{1/n} | \mathcal{F}_0) - \xi/n) \right|$$

$$\leq \frac{M}{n} \sum_0^{n-1} E \, | n E(\alpha^*_{1/n} | \mathcal{F}_0) - \xi |$$

$$= M \| n E(\alpha^*_{1/n} | \mathcal{F}_0) - \xi \| \to 0 \quad \text{by assumption.}$$

Putting (40), (41) and (42) together, we finally obtain (38).

REFERENCES

1. R. M. Blumenthal and R. K. Getoor, *Markov processes and potential theory*, Pure and Appl. Math., vol. 29, Academic Press, New York and London, 1968. MR 41 #9348.

2. C. Dellacherie, *Capacités et processus stochastiques*, Springer-Verlag, Berlin and New York, 1972.

3. H. Föllmer, *On the representation of semimartingales*, Ann. Probability 1 (1973), 580–589.

4. D. Geman and J. Horowitz, *Occupation times for smooth stationary processes*, Ann. Probability 1 (1973), 131–137.

5. ———, *Remarks on Palm measures*, Ann. Inst. H. Poincaré 9 (1973), 215–232.

6. ———, *Random shifts which preserve measure*, Proc. Amer. Math. Soc. 49 (1975), 143–150.

7. A. Hanen, *Processus ponctuels stationnaires et flots spéciaux*, Ann. Inst. H. Poincaré Sect. B 7 (1971), 23–30. MR 46 #2727.

8. J. de Sam Lazaro and P. A. Meyer, *Méthodes des martingales et théorie des flots*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 18 (1971), 116–140.

9. ———, *Questions de théorie des flots*, Université de Strasbourg, Séminaire de Probabilités.

10. B. Maisonneuve, *Ensembles régénératifs, temps locaux et subordinateurs*, Lecture Notes in Math., no. 191, Springer-Verlag, Berlin and New York, 1971.

11. J. Mecke, *Stationäre zufällige Masse auf lokalkompakten Abelschen Gruppen*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 9 (1967), 36–58. MR 37 #3611.

12. ———, *Invarianzeigenschaften allgemeiner Palmsche Masse* (to appear).

13. P. A. Meyer, *Probability and potentials*, Blaisdell, Waltham, Mass., 1966. MR 34 #5119.

14. F. Papangelou, *Integrability of expected increments of point processes and a related random change of scale*, Trans. Amer. Math. Soc. 165 (1972), 483–507. MR 47 #2654.

15. K. R. Parthasarathy, *Probability measures on metric spaces*, Probability and Math. Statist., no. 3, Academic Press, New York, 1967. MR 37 #2271.

16. D. Revuz, *Mesures associées aux fonctionnelles additives de Markov*. I, Trans. Amer. Math. Soc. 148 (1970), 501–531. MR 43 #5611.

17. H. Totoki, *Time changes of flows*, Mem. Fac. Sci. Kyushu Univ. Ser. A 20 (1966) 27–55. MR 34 #1488.

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF MASSACHUSETTS, AMHERST, MASSACHUSETTS 01002

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

GROUP PRESENTATIONS AND FORMAL DEFORMATIONS([1])

BY

PERRIN WRIGHT

ABSTRACT. Formal deformations (expansions and collapses) of dimension ≤ 3 among 2-dimensional polyhedra are explained in terms of a certain collection of operations on finite group presentations. The results are valid for any simple homotopy type of 2-dimensional polyhedra, and simplifications are possible within the simply connected simple homotopy types.

1. Introduction. The relationship between finite group presentations and finite 2-dimensional polyhedra is in evidence at various places in the literature. Furthermore, the folklore has it that there exists a correspondence from the category of finite group presentations and certain operations thereon, to the category of 2-polyhedra and 3-dimensional formal deformations (expansions and collapses). The purpose of this paper is to give a precise formulation of the problem, with solutions, thereby exonerating the folk. The references listed here, with the exception of [2], are articles in which this problem has been addressed to some extent.

Whitehead showed [5] that any two n-polyhedra having the same simple homotopy type are formally equivalent under deformations of dimension $n+1$, provided $n > 2$. For $n = 2$, one must apparently deform through 4-dimensional polyhedra. The reducibility of the dimension of the deformation to three is equivalent to the group theoretic problem to be described here.

The author is greatly indebted to the referee for suggesting a proof of Lemma 2 and for contributing substantially to §8.

2. The complexes. We shall initially restrict our attention to a special class C of 3-dimensional CW-complexes, defined as follows: If $X \in C$, then:

(1) $X^{(0)}$ consists of a single 0-cell v.

(2) $X^{(1)}$ is the union of X^0 and a finite collection $\{x_i\}$ of 1-cells whose boundaries are attached to v.

(3) $X^{(2)}$ is obtained from $X^{(1)}$ by attaching to $X^{(1)}$ a finite collection

Received by the editors December 20, 1973 and, in revised form, April 17, 1974.
AMS (MOS) subject classifications (1970). Primary 57C10; Secondary 20F05.
([1])This research was supported by COFRS, Florida State University.

$\{e_i\}$ of 2-cells, where Bd e_k is subdivided into edges and vertices; each vertex of Bd e_i is sent to v and each (open) edge of Bd e_i is sent either to v or homeomorphically to some (open) x_i.

(4) $X^{(3)}$ is obtained by attaching to $X^{(2)}$ a finite collection $\{d_i\}$ of 3-cells, where Bd d_i has the structure of a cell complex; each vertex of Bd d_i is sent to v, each edge to v or homeomorphically to some x_i, and each (open) 2-cell homeomorphically to some (open) e_i, subject to the usual condition that the attaching map be continuous.

3. Elementary expansions and collapses in C. An *elementary n-expansion* $K \nearrow L$ in C is defined provided $L = K \cup_f B^n$, where f attaches (as in 2(4)) to K all of the boundary of B^n except one (open) $(n-1)$-cell. An *elementary n-collapse* in C is the inverse of an elementary n-expansion, written $L \searrow K$.

There are no 1-expansions or 1-collapses in C because each $K \in C$ has only one vertex.

A *formal n-deformation* from K to L in C is a finite sequence $\{K_0, \ldots, K_m\}$ in C such that $K_0 = K$, $K_m = L$, K_i expands or collapses elementarily to K_{i+1}, and dim $K_i \leq n$ for all i.

4. Presentations. We depart somewhat from the usual definition of a presentation in order to obtain a correspondence between presentations and complexes in C. A finite group presentation will herein consist of a set $\{x_i\}$ of distinct symbols, called the generators, together with a set $\{r_i\}$ of distinct symbols, called the relators; $\{x_i\}$ and $\{r_i\}$ shall be indexed by finite subsets of the natural numbers. Associated with each relator r_i is a word ρ_i (not necessarily reduced) in the generators $\{x_i\}$, and the group presented is the quotient group $F\{x_i\}$ modulo the normal closure of the $\{\rho_i\}$. We shall use the standard notation $\{\{x_1\}|\{r_i\}\}$ for a presentation.

Two presentations $\{\{x_i\}|\{r_i\}\}$ and $\{\{y_i\}|\{s_i\}\}$ will be considered equal if and only if there exist 1-1 correspondences $\{x_i\} \leftrightarrow \{y_i\}$ and $\{r_i\} \leftrightarrow \{s_i\}$ which preserve the words associated with the relators.

Associated with each presentation $p = \{\{x_i\}|\{r_i\}\}$ is a 2-complex $K(p) \in C$, unique up to homeomorphism, which is obtained by attaching 1-cells $\{x_i\}$ to v, then attaching 2-cells $\{e_i\}$ along their boundaries by the words $\{\rho_i\}$. (If ρ_i is the empty word $\emptyset$, then ∂e_i is attached to v.) Then $\pi_1 K(p)$ is the group presented by p.

Conversely, if K is an oriented 2-complex in C, a group presentation $p(K)$ is induced, which is unique up to indexing of the $\{x_i\}$ and $\{r_i\}$ and cyclic permutation of the $\{\rho_i\}$.

Remark 1. If $i \neq j$, then r_i and r_j are distinct relators, even if $\rho_i = \rho_j$. For example, let $p_1 = \{x_1 | r_1\}$ and $p_2 = \{x_1 | s_1, s_2\}$, where the words associated with r_1, s_1, s_2 are all x_1. Then $p_1 \neq p_2$, and $K(p_1)$ is a 2-cell while $K(p_2)$ is a 2-sphere. We shall generally abuse our notation when no ambiguity is present, and write $p_2 = \{x_1 | x_1, x_1\}$; that is, we shall use ρ_i instead of r_i in describing the presentation, suppressing (but not forgetting) the indexing of the relators.

Remark 2. Cancellation of adjacent inverses within a relator, taken for granted in word operations, will not be allowed here. For example, if $p_1 = \{x | \emptyset\}$ and $p_2 = \{x | xx^{-1}\}$, then $K(p_1) = S^1 \vee S^2$ but $K(p_2)$ is a pinched S^2. We shall show later that cancellation corresponds to a formal 3-deformation.

Remark 3. Each relator word ρ is assumed to be written as a noncollected word in the generators and their inverses; that is, $\rho = x_1^2 x_2^3$ should be written $x_1 x_1 x_2 x_2 x_2$. In constructing $K(p)$, the corresponding 2-cell may have some boundary edges which are identified to v, as long as the remaining edges, taken clockwise from some point, read the word ρ. The insertion of edges to be identified with v does not change the homeomorphism type of $K(p)$ and corresponds to insertion of the identity element of $F(x_1, \ldots, x_n)$ at various places within the relator word ρ.

5. **2-dimensional operations.** On a presentation $p = \{x_1, \ldots, x_n | r_1, \ldots, r_k\}$, define the following operations:

(1) Cyclically permute the letters of any ρ_i.

(2) Replace ρ_i by ρ_i^{-1}.

(3) Add a generator α and a relator (whose word is) αw, where w is a word in $x_1, \ldots, x_n$ (possibly $\emptyset$).

(4) Delete a generator α and a relator αw, provided that α does not appear in any other relator or in w.

Of these operations, only (3) and (4) alter the homeomorphism type of $K(p)$.

Theorem 1. K^2 *formally 2-deforms to* L^2 *in* C *if and only if* $p(K)$ *can be transformed to* $p(L)$ *by operations* (1), (2), (3), (4).

Proof. Since there are no 1-deformations in C, it suffices to consider a single elementary 2-expansion or collapse. A 2-expansion $K \nearrow L$ consists of adding a 1-cell α and a 2-cell e whose boundary is attached via the word αw, where w is any word in the 1-cells of K. By performing (3) on $p(K)$, followed by (1), (2), if necessary, we obtain $p(L)$. A 2-collapse $K \searrow L$ corresponds to (4); there must be a free edge α through which to collapse

a 2-cell e. In $p(K)$ a appears once in the relator r corresponding to e, and in no other relator. Say $r = waw'$. Apply (1) to get $aw'w$, then (4) to delete the generator a and relator $aw'w$. Follow with (1), (2) if necessary.

Conversely, each operation on a presentation p may be realized on $K(p)$ as follows: for (1), (2), do nothing to $K(p)$; for (3), expand; for (4), collapse.

6. **The 3-dimensional operation.** A 3-deformation between K^2 and L^2 in C will be called *transient* if each 3-expansion is followed immediately by a 3-collapse. There is no accumulation of 3-cells in a transient deformation. Lemma 2 shows that we need devise a presentation operation for transient 3-deformations only.

Lemma 2. *If K^2 3-deforms to L^2 in C, then K^2 transiently 3-deforms to L^2 in C.*

Proof. Let a 3-deformation D be given. Enumerate the 3-cells $E_1, \ldots, E_n$ in the order in which they appear in D, and let F_i denote the face through which E_i is eventually collapsed.

Construct a transient deformation D' in the following manner. When E_1 is attached in D, let it be attached in D' but immediately collapsed via F_1. When E_2 is attached in D, let it be attached in D' such that any faces which were attached to F_1 in D are now subdivided and attached to $\partial E_1 - \mathring{F}_1$ instead, via some map ϕ_1 induced by $F_1 \subset E_1 \searrow \partial E_1 - \mathring{F}_1$. The face F_2 must now be free: this fails only if, in D, $F_1 \subset \partial E_2$ and $F_2 \subset \partial E_1$, which is impossible since it would block the collapse of both E_1 and E_2 in D. Collapse E_2 via F_2.

In a similar fashion, let each subsequent 3-cell E_i be attached in D' via the composition of its attaching map in D and the maps $\phi_{i-1}, \ldots, \phi_1$, then collapsed immediately via F_i, which must be free or else there would exist a circle of inequalities $F_i \subset \partial E_{k_1}, F_{k_1} \subset \partial E_{k_2}, \ldots, F_{k_m} \subset \partial E_i$, blocking the collapses of $E_i, E_{k_1}, \ldots, E_{k_m}$ in D.

We shall now describe an operation on a presentation $\{x_i | r_i\}$ which corresponds to an elementary transient 3-deformation $K^2 \nearrow H^3 \searrow L^2$.

A word ρ in $x_1, \ldots, x_n$ will be called *allowable* if it is obtained by the following steps:

(0) Beginning with the empty word, successively insert words of the form xx^{-1} or $x^{-1}x$ at any point in the word, where x is any generator. Call this word s.

(1) Choose any relator r_i, and let ρ_i' be any cyclic permutation of ρ_i. Let s' be any cyclic permutation of s.

(2) Form the product $s'\rho_i'$.

(3) Optionally perform any cancellations induced by juxtaposition of s' and ρ_i'.

(4) Call the new word s again, and iterate steps (1), (2), (3).

Operation (5). If ρ is an allowable word in $\{x_i\}$ and if r_* is some relator which is used exactly once in constructing ρ, change ρ_* to ρ^{-1}.

To see that (5) corresponds to an elementary transient 3-deformation, list the relators in the order in which they were used in constructing ρ, say $r_{i_1}, \ldots, r_{i_n}$.

In S^2, construct a tree t whose edges read counterclockwise (from $S^2 - t$) the word s of step (0). Construct a 2-cell e_{i_1}, whose boundary edges read ρ_{i_1}' and whose intersection with t corresponds to the cancellations (if any) in step (3), so that the word read counterclockwise from $S^2 - (t \cup e_{i_1})$ is the word s of step (4). (If $\rho_{i_1} = \emptyset$, let e_{i_1} have one boundary edge labelled v.)

Modelling on operation (5) we construct in this fashion a collapsible cell complex $t \cup e_{i_1} \cup \ldots \cup e_{i_r}$ in S^2, and the word read counterclockwise from the complement is the word ρ. Let e denote the complementary 2-cell; its clockwise boundary word is ρ^{-1}.

Attach B^3 to $K(p)$ using this subdivision of S^2, by mapping each vertex of S^2 to v, each edge to the 1-cell or vertex of $K(p)$ whose letter it bears, and each open 2-cell e_{i_k} homeomorphically to its counterpart in $K(p)$. This move is an elementary expansion, of which e is the free face.

Let e_* be the 2-cell of $K(p)$ corresponding to r_*. Since r_* was used exactly once in constructing ρ, e_* is a free face of the new 3-cell. Collapse the 3-cell via e_*. This transient 3-deformation realizes operation (5).

Conversely, consider a 3-deformation $K \nearrow K \cup_f B^3 \searrow L$. Let $p(K) = \{\{x_i\}|\{r_i\}\}$. The expansion is accomplished by attaching $S^2 - \mathring{e}$ to K, where e is some 2-cell in some cell subdivision of S^2. Let ρ^{-1} be the word read clockwise from Bd e. Then $S^2 - \mathring{e}$ is a collapsible complex whose boundary word (read counterclockwise from e) is ρ. Collapse the 2-cells of $S^2 - \mathring{e}$ in any order and let t be the remaining tree. Each edge of t is mapped by f to some 1-cell x_i of K. Expand from any vertex of t to t itself; this induces step (0) of operation (5). The 2-expansion $t \nearrow S^2 - \mathring{e}$ induces steps (1), (2), (3) for each 2-cell in the expansion. As a result, the word ρ is

built up in an allowable fashion from the presentation $p(K)$. If e_* is the free face in the collapse $K \cup_f B^3 \searrow L$, it must follow that r_* was used exactly once in constructing ρ. Apply operation (5) to replace the relator word ρ_* by ρ^{-1}. This operation realizes the transient 3-deformation.

The following theorem has now been established.

Theorem 2. *K^2 formally 3-deforms to L^2 in C if and only if $p(K)$ can be transformed to $p(L)$ by operations* (1) *through* (5).

7. **Consequences of the operations.** The following operations can be performed as a composition of operations (1)–(5).

Cancellation. Suppose some relator r has associated word $\rho = uxx^{-1}v$, where u and v are words in $\{x_i\}$. We may replace ρ by uv by the operations

$$\begin{aligned}\{|uxx^{-1}v\} &\xrightarrow{1} \{|xx^{-1}w\} \xrightarrow{3} \{a|xx^{-1}w,\ ax^{-1}\}\\ &\xrightarrow{5} \{a|w^{-1}xa^{-1},\ ax^{-1}\} \xrightarrow{5} \{a|w,\ ax^{-1}\}\\ &\xrightarrow{4} \{|w\} \xrightarrow{1} \{|uv\}.\end{aligned}$$

Conjugation. To replace ρ by $g^{-1}\rho g$, where g is any word in $\{x_i\}$, do repeated applications of the sequence $\rho \to xx^{-1}\rho \to x^{-1}\rho x$.

Forming products. If r_i, r_j are relators and $i \neq j$, we may replace ρ_i by $\rho_i\rho_j$ by

$$\{|\rho_i, \rho_j\} \xrightarrow{5} \{|(\rho_i\rho_j)^{-1}, \rho_j\} \xrightarrow{2} \{|\rho_i\rho_j, \rho_j\}.$$

It is necessary for operation (5) that $i \neq j$, to ensure that r_i is used exactly once in constructing $\rho_i\rho_j$.

8. **Generalization to polyhedra.** It is desirable to generalize Theorem 2 to a theorem about polyhedra. The necessary ingredients are the representation of polyhedra by elements of C, and the invariance of 3-deformation classes under this representation.

The first generalization is to cell complexes. If K is any cell complex and T is any tree in K which contains all vertices, then $K/T \in C$.

Lemma 3. *K^2 formally 3-deforms (through cell complexes) to K/T. If T_0, T_1 are trees in K, then K/T_0 3-deforms in C to K/T_1.*

Proof. Let $K \times I$ have the product structure. Then $(K \times I)/(T \times 1) \searrow (K \times 1)/(T \times 1) \cong K/T$. Also

$$(K \times I)/(T \times 1) \searrow (K \times 0) \cup (T \times I)/(T \times 1) \searrow K \times 0,$$

since $T \searrow 0$. Thus K 3-deforms to K/T.

Let v be any vertex of K. Let $T = (T_0 \times 0) \cup (v \times I) \cup (T_1 \times 1)$. Then $K \times I \searrow (K \times 0) \cup T$ by collapsing $K \times I$ vertically to $(K \times 0) \cup (T_1 \times I)$, then collapsing $T_1 \times I$ horizontally to $(T_1 \times 0) \cup (T_1 \times 1) \cup (v \times I)$. Upon smashing T, we obtain $(K \times I)/T \searrow (K \times 0)/T_0 \cong K/T_0$. Similarly $(K \times I)/T \searrow (K \times 1)/T_1 \cong K/T_1$. Since $(K \times I)/T$ has one vertex, K/T_0 3-deforms in C to K/T_1.

Lemma 4. *Let K^2 and L^2 be cell complexes and let K^2 3-deform cellularly to L^2. Let T and U be trees in K and L which contain all vertices. Then K/T 3-deforms in C to L/U.*

Proof. There is a cell complex H^3 such that $K^2 \nearrow H^3 \searrow L^2$. The complex H^3 is obtained by reordering the 3-deformation from K to L so that all expansions occur first.

Let $T_1 = T \cup$ (trail of vertices in $H^3 \searrow K^2$) and $U_1 = U \cup$ (trail of vertices in $H^3 \searrow L^2$). Now T_1 contains no free edges of $H \searrow K$, so this collapse induces a collapse $H/T_1 \searrow K/T$ in C.

Let X^2 be the 2-complex which remains after collapsing the 3-cells in $H \searrow L$. Then $T_1 \subset X$, and $H/T_1 \searrow X/T_1$ in C. By Lemma 3, X/T_1 3-deforms in C to X/U_1, which in turn collapses to L/U in C.

Lemma 5. *Let K^2 be a cell complex and let K' be a cell subdivision of K. Let T and T' be trees in K and K' containing all vertices. Then K/T 3-deforms to K'/T' in C.*

Proof. Let $K \times I$ have the product structure induced from K, except on $K \times 1$ where the structure is induced from K'. Then $K \cong K \times 0 \nearrow K \times I \searrow K \times 1 \cong K'$ as cell complexes, and by Lemma 4, K/T 3-deforms to K'/T' in C.

For an arbitrary compact connected 2-polyhedron P, a *representative $\overline{P}$ of P in C* is obtained by triangulating P in any fashion as a cell complex and smashing any tree containing all vertices. A *presentation induced by P* is any presentation $p(\overline{P})$, where $\overline{P}$ is a representative of P in C.

Theorem 3. *The following are equivalent:*

(i) *The polyhedron P^2 formally 3-deforms (polyhedrally) to the polyhedron Q^2.*

(ii) *For some representatives $\overline{P}$, $\overline{Q}$ in C, $\overline{P}$ 3-deforms to $\overline{Q}$ in C.*

(iii) *For all representatives $\overline{P}$, $\overline{Q}$ in C, $\overline{P}$ 3-deforms to $\overline{Q}$ in C.*

By virtue of Theorem 2, we obtain

Corollary 3.1. *The polyhedron P^2 formally 3-deforms to the polyhedron Q^2 if and only if some (all) induced presentation(s) of P can be transformed to some (all) induced presentation(s) of Q by operations* (1)–(5).

Proof of Theorem 3. (iii)$\rightarrow$(i). Let $\overline{P}$, $\overline{Q}$ be representatives of P, Q in C. By Lemma 3, there exist (polyhedral) 3-deformations $P \rightarrow \overline{P}$, $Q \rightarrow \overline{Q}$, and by hypothesis, $\overline{P}$ 3-deforms to $\overline{Q}$. Hence P 3-deforms polyhedrally to Q.

(i)$\rightarrow$(ii). If P^2 3-deforms to Q^2, there exists a polyhedron Z^3 such that $P \nearrow Z \searrow Q$. There exist simplicial triangulations H, K, L of Z, P, Q such that $K \nearrow H \searrow L$ simplicially. (These may be obtained by triangulating Z, subdividing to get a simplicial collapse to P, subdividing further to get a simplicial collapse to Q, and invoking [2] to see that the simplicial collapse to P is not lost.) Since K 3-deforms to L simplicially, Lemma 4 states that for any representatives $\overline{P}$, $\overline{Q}$ of the form K/T, L/U (*for these particular* K, L), $\overline{P}$ 3-deforms to $\overline{Q}$ in C.

(ii)$\rightarrow$(iii). If K/T and K'/T' are representatives of P in C, then K and K' have a common (up to isomorphism) subdivision K''. Let T'' be any tree in K'' containing all vertices. Then by Lemma 5, there are 3-deformations $K/T \rightarrow K''/T'' \rightarrow K'/T'$ in C. Hence if $\overline{P}$ 3-deforms to $\overline{Q}$ in C for some representatives, the same is true for all representatives.

9. Simply connected complexes. If $K \in C$ has $\pi_1(K) = 1$ then in $p(K) = \{x_1, \dots, x_n | r_1, \dots, r_k\}$, the normal closure of $r_1, \dots, r_k$ in the free group $F(x_1, \dots, x_n)$ is the free group $F(x_1, \dots, x_n)$. With this extra condition, the operations (1)–(5) can be simplified to these:

(0) Cancellation (and its inverse).

(i) Replace r_i by r_i^{-1}.

(ii) Replace r_i by $r_i r_j$, $i \neq j$.

(iii) Replace r_i by $g^{-1} r_i g$, $g \in F(x_1, \dots, x_n)$.

(iv) Add a generator x and a relator x.

(v) Delete a generator x and relator x if x appears in no other relator.

It is easily seen that operations (1)–(5) imply the new operations. The converse is also true, and only (3) and (5) present any difficulty. To obtain (3), write w as a product of conjugates of the relators, then add $\{a | a\}$ and apply (ii) and (iii) repeatedly to change the relator a to aw. To obtain (5), let w be the word built in the process of constructing ρ just prior to the usage of the relator r_*. Since w is a product of conjugates of the other relators, we can replace ρ_* by $w\rho_*$. Then ρ is constructed without

using r_* again, so we may replace $w\rho_*$ by ρ, then ρ^{-1} to get (5).

From the foregoing and Corollary 3.1 we have

Corollary 3.2. *P^2 formally 3-deforms to an n-fold wedge of 2-spheres if and only if all presentations induced by P can be transformed to the presentation with no generators and n empty relators by the operations* (0), (i), . . . , (v).

When $n = 0$, this says that contractible 2-polyhedra 3-deform to a point if and only if their induced presentations can be transformed to the empty presentation $\{\ |\ \}$ by those operations.

REFERENCES

1. J. J. Andrews and M. L. Curtis, *Free groups and handlebodies*, Proc. Amer. Math. Soc. 16 (1965), 192–195. MR 30 # 3454.

2. D. R. J. Chillingworth, *Collapsing three-dimensional convex polyhedra*, Proc. Cambridge Philos. Soc. 63 (1967), 353–357. MR 35 # 995.

3. H. F. Trotter, *Homology of group systems with applications to knot theory*, Ann. of Math. (2) 76 (1962), 464–498. MR 26 # 761.

4. C. T. C. Wall, *Formal deformations*, Proc. London Math. Soc. (3) 16 (1966), 342–352. MR 33 # 1851.

5. J. H. C. Whitehead, *Simplicial spaces, nuclei, and m-groups*, Proc. London Math. Soc. (2) 45 (1939), 243–327.

6. ———, *On incidence matrices, nuclei, and homotopy types*, Ann. of Math. (2) 42 (1941), 1197–1239. MR 3, 142.

DEPARTMENT OF MATHEMATICS, FLORIDA STATE UNIVERSITY, TALLAHASSEE, FLORIDA 32306

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

CONVERGENT SUBSEQUENCES FROM SEQUENCES OF FUNCTIONS(1)

BY

JAMES L. THORNBURG

ABSTRACT. Let $\{y_k\}$ be a sequence of functions, $y_k \in \Pi_{s\in S} E_s$ where S is a nonempty subset of the l-dimensional Euclidean space and E_s is an ordered vector space with positive cone K_s. If $y_k \in \Pi_{s\in S} E_s$, sufficient conditions are given that $\{y_k\}$ have a subsequence $\{h_k\}$ such that for each $t \in S$ the sequence $\{h_k(t)\}$ is monotone for k sufficiently large, depending on t. If each E_s is an ordered topological vector space, sufficient conditions are given that $\{y_k\}$ has a subsequence $\{h_k\}$ such that for every $t \in S$ the sequence $\{h_k(t)\}$ is either monotone for k sufficiently large depending on t, or else the sequence $\{h_k(t)\}$ is convergent. If $E_s = B$ for each s and B a Banach space then a definition of bounded variation is given so that if $\{y_k\}$ is uniformly norm bounded and the variation of the functions y_k is uniformly bounded then there is a convergent subsequence $\{h_k\}$ of $\{y_k\}$. In the case $E_s = E$ for each $s \in S$ and E is such that bounded monotone sequences converge then the given conditions imply the existence of a subsequence $\{h_k\}$ of $\{y_k\}$ which converges for each $t \in S$.

1. Introduction. The terminology used in this paper relating to ordered vector spaces will agree with that of Peressini [10] unless stated otherwise or explicitly defined. In particular, the definitions of ordered vector space E, positive cone K of E, ordered interval $[a, b]$ in E where $a \leq b$, ordered bounded subset of E, majorized subset of E, ordered topological vector space E, ordered locally convex space E, and normal positive cone K of a topological vector space E agree with [10]. An ordered Banach space B will be a Banach space which is also an ordered vector space and thus is an ordered topological vector space. We note that we do not require the positive cone K of a topological vector space E to be closed as is done, for example, in

Received by the editors April 19, 1974.

AMS (MOS) subject classifications (1970). Primary 40A05, 26A45; Secondary 26A46.

Key words and phrases. Sequences of functions, convergence of subsequences, monotonicity of subsequences, variation, functions of bounded variation, ordered topological vector spaces.

(1) This work is in partial fulfillment of the requirements for the Doctor of Philosophy degree at the University of Missouri, Columbia.

Shaefer [16]. The l-dimensional Euclidean space R^l will be considered with the usual norm unless otherwise stated, the open sphere will be $D(t, \delta) = \{x \in R^l: \|t - x\| < \delta\}$ with center t and radius δ and the order will be determined by some positive cone

$$Q_k = \{(x^{(1)}, x^{(2)}, \dots, x^{(l)}): x^{(l)} \geq 0, x^{(i)} \geq 0 \text{ for } t_i = 0, x^{(i)} \leq 0 \text{ for } t_i = 1 \text{ where } k = 1 + t_1 + 2t_2 + 2^2 t_3 + \cdots + 2^{(l-2)} t_{l-1}\},$$
$$1 \leq k \leq 2^{l-1}.$$

The interior of the positive cone will be

$$Q_k^\circ = \{(x^{(1)}, x^{(2)}, \dots, x^{(l)}): x^{(l)} > 0, x^{(i)} > 0 \text{ if } t_i = 0, x^{(i)} < 0 \text{ if } t_i = 1 \text{ for } k = 1 + t_1 + 2t_2 + 2^2 t_3 + \cdots + 2^{(l-2)} t_{l-1}\}.$$

The translate of any $l - 1$ dimensional subspace will be called a hyperplane in R^l and if it is of the form $\{(\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(l)}): \alpha^{(i)} = c$ for i fixed and some constant $c\}$ it will be called a regular hyperplane. A closed interval will be

$$I = [\alpha, \beta] = \{(x^{(1)}, x^{(2)}, \dots, x^{(l)}): \alpha^{(i)} \leq x^{(i)} \leq \beta^{(i)}, i = 1, 2, \dots, l\}$$

where $\alpha = (\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(l)})$, $\beta = (\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(l)})$ and $\alpha^{(i)} < \beta^{(i)}$ for $i = 1, 2, \dots, l$. The vertices of I will be $\{x_1, x_2, \dots, x_{2^l}\}$ where

$$x_k = (y^{(1)}, y^{(2)}, \dots, y^{(l)}), \quad y^{(i)} = \begin{cases} \alpha^{(i)} & \text{if } t_i = 0, \\ \beta^{(i)} & \text{if } t_i = 1, \end{cases} \quad i = 1, 2, \dots, l,$$

and

$$k = 1 + t_1 + 2t_2 + \cdots + 2^{(l-1)} t_l.$$

The set $\{(y^{(1)}, y^{(2)}, \dots, y^{(l)}): y^{(m)} = \alpha^{(m)}$ for m fixed and $\alpha^{(i)} \leq y^{(i)} \leq \beta^{(i)}$ for $i \neq m\}$ will be called a minimum bounding edge of I and $\{(y^{(1)}, y^{(2)}, \dots, y^{(l)}): y^{(m)} = \beta^{(m)}$ for m fixed and $\alpha^{(i)} \leq y^{(i)} \leq \beta^{(i)}$ for $i \neq m\}$ will be called a maximum bounding edge of I.

Two elements x_k and x_j in an ordered vector space E with positive cone K will be comparable if $x_k - x_j \in K$ or $x_j - x_k \in K$ holds. By a monotone nondecreasing sequence $\{x_k\}$ in an ordered vector space E we mean $x_{k+1} - x_k \in K$ holds for all k. A monotone nonincreasing sequence is similarly defined and a sequence is called monotone if it is either monotone nondecreasing ot monotone nonincreasing.

A sequence $\{x_k\}$ in an ordered vector space E will be said to be eventually monotone nondecreasing if there exists an integer k_0 so that $x_{k+1} - x_k \in K$ holds for all $k \geq k_0$. An eventually monotone nondecreasing sequence is similarly defined and a sequence is called eventually monotone if it is either eventually monotone nondecreasing or eventually monotone nonincreasing.

If $\{y_k\}$ is a sequence of functions, $y_k \in \Pi_{s \in S} E_s$ where S is a nonempty subset of R^l and E_s is an ordered vector space with positive cone K_s then we say that the sequence $\{y_k\}$ is a monotone nondecreasing sequence of functions on S if the sequence $\{y_k(s)\}$ is a monotone nondecreasing sequence in E_s for each $s \in S$. A monotone nonincreasing sequence of functions on S is similarly defined and a sequence of functions is said to be monotone on S if it is either monotone nondecreasing or monotone nonincreasing.

If $\{y_k\}$ is a sequence of functions, $y_k \in \Pi_{s \in S} E_s$, where S is a nonempty subset of R^l and E_s is an ordered vector space with positive cone K_s then we say that the sequence of functions $\{y_k\}$ is an eventually monotone nondecreasing sequence on S if $\{y_k(s)\}$ is eventually monotone for each $s \in S$.

Let $\{y_k\}$ be a sequence of functions, $y_k \in \Pi_{s \in S} E_s$ where S is a nonempty subset of R^l and each E_s is an ordered vector space. Sufficient conditions that $\{y_k\}$ have a subsequence $\{h_k\}$ such that for each $s \in S$, $\{h_k(s)\}$ is eventually monotone in E_s are given in Theorem 4.1. In case each E_s is a sequentially complete ordered locally convex space, Theorem 4.4 gives sufficient conditions that there be a subsequence $\{h_k\}$ of $\{y_k\}$ such that for every $s \in S$ the sequence $\{h_k(s)\}$ is either eventually monotone or else convergent in E_s. Additional hypotheses are given in Theorem 5.1 to insure the existence of a subsequence $\{h_k\}$ of $\{y_k\}$ such that $\{h_k(s)\}$ converges in E_s for each $s \in S$. This generalizes the results in [15] and gives, in Theorem 5.4, a necessary and sufficient condition that a sequence of functions from a subset of l-dimensional Euclidean space into q-dimensional Euclidean space have a subsequence that is pointwise convergent.

For $\{y_k\}$ a sequence of functions, $y_k: S \to B$ where S is a nonempty subset of R^l and B is an ordered Banach space, a definition of variation is given in §3 so that Theorem 5.2 yields a generalized Helly selection theorem as a corollary.

2. Preliminary results. We will refer to a result of Ramsey [11, Theorem A, p. 264] or [12, Theorem A, p. 82] repeatedly so we will include its statement here.

Theorem 2.1. *Let Γ be an infinite class, u and r positive integers, and let those subclasses of Γ which have exactly r members, or, as we may*

say, let all r-combinations of the members of Γ *be divided in any manner into* u *mutually exclusive classes* C_i $(i = 1, 2, \ldots, u)$, *so that every r-combination is a member of one and only one* C_i. *Then, assuming the axiom of selections,* Γ *must contain an infinite subclass* Δ *such that all the r-combinations of the members of* Δ *belong to the same* C_i.

Corollary 2.2. *Let* J *be a nonempty subset of* R^l *and* $\{f_k\}$ *be a sequence of functions,* $f_k \in \Pi_{t \in J} E_t$ *where each* E_t *is an ordered vector space with positive cone* K_t. *If* $f_k(t)$ *and* $f_j(t)$ *are comparable for each* $t \in J$ *and* $k, j = 1, 2, \ldots$, *then either there is a subsequence* $\{h_j\}$ *of* $\{f_k\}$ *such that* $\{h_j\}$ *is a monotone subsequence on* J *or else there is a subsequence* $\{h_j\}$ *of* $\{f_k\}$ *such that if* $i \neq j$ *there are* $t, \tau \in J$ *depending on* i, j *with* $h_i(t) - h_j(t) \in K_t$, $h_i(t) \neq h_j(t)$ *and* $h_j(\tau) - h_i(\tau) \in K_\tau$, $h_j(\tau) \neq h_i(\tau)$.

The proof of this corollary follows in a similar manner to that of Corollary 2.3 in [14].

Corollary 2.3. *Let* J *be a nonempty subset of* R^l *and* $\{f_k\}$ *be a sequence of functions,* $f_k \in \Pi_{t \in J} E_t$, *where each* E_t *is an ordered topological vector space. If* W_t *is a circled neighborhood of* θ_t *in* E_t *then either there is a subsequence* $\{h_j\}$ *of* $\{f_k\}$ *such that for* $i \neq j$, $h_j(t) - h_i(t) \in W_t$ *for all* $t \in J$ *or else there is a subsequence* $\{h_j\}$ *of* $\{f_k\}$ *such that for* $i \neq j$, $h_j(t) - h_i(t) \notin W_t$ *for some* $t \in J$ *depending on* i, j.

The proof of this corollary follows in a similar manner to that of Corollary 2.3 in [14].

3. **Variation.** In [14] where sequences of functions from the real numbers into the real numbers were considered, a result yielding the Helly selection theorem as a corollary was given. Then in [15] where the functions were from the real numbers into ordered Banach spaces, a generalized Helly selection theorem was given as a corollary. Browne [1] considers functions from l-dimensional Euclidean space into the real numbers and with one definition of variation gets another generalized Helly selection theorem. Four definitions of variation are given here and their relationship considered when the functions are from l-dimensional Euclidean space into an ordered Banach space.

Consider the interval $I = [\alpha, \beta] \subset R^l$ and $a, b \in I$ with $a^{(i)} < b^{(i)}$ for $i = 1, 2, \ldots, l$, then $J = [a, b]$ will be a subinterval of I. A subdivision of $[\alpha, \beta]$ will consist of a finite number of subintervals $J_1, J_2, \ldots, J_m$ of I such that $\bigcup_{i=1}^m J_i = I$ and for $k \neq j$, $J_k \cap J_j$ is either empty, a point, or is a p-dimensional interval where $1 \leq p \leq l - 1$. A refinement of the subdivision $\{J_1, J_2, \ldots, J_m\}$ will be a subdivision obtained by subdividing one or more

of the J_i's into two or more subintervals. A subdivision $\{J_1, J_2, \ldots, J_m\}$ will be called regular if all of the minimum bounding edges of $J_n = [a_n, b_n]$,

$$\{(y^{(1)}, y^{(2)}, \ldots, y^{(l)})\colon y^{(j)} = a_n^{(j)},\ a^{(i)} \le y^{(i)} \le b^{(i)} \text{ for } i \ne j\}$$

and the maximum bounding edges of J_n,

$$\{(y^{(1)}, y^{(2)}, \ldots, y^{(l)})\colon y^{(j)} = b_n^{(j)},\ a_n^{(i)} \le y^{(i)} \le b^{(i)} \text{ for } i \ne j\}$$

for $n = 1, 2, \ldots, m$ are extended to

$$\{(y^{(1)}, y^{(2)}, \ldots, y^{(l)})\colon y^{(j)} = a_n^{(j)},\ \alpha^{(i)} \le y^{(i)} \le \beta^{(i)} \text{ for } i \ne j\}$$

and

$$\{(y^{(1)}, y^{(2)}, \ldots, y^{(l)})\colon y^{(j)} = b_n^{(j)},\ \alpha^{(i)} \le y^{(i)} \le \beta^{(i)} \text{ for } i \ne j\},$$

respectively, and the subdivision is not refined. A set $X = \{x_1 = \alpha, x_2, x_3, \ldots, x_{n-1}, x_n = \beta\} \subset I = [\alpha, \beta]$ will be called a partition of I. The subdivision whose bounding edges are formed by the intersection of I with all the regular hyperplanes through $\{x_1, x_2, \ldots, x_n\}$ will be called the subdivision determined by the partition $\{x_1, x_2, \ldots, x_n\}$. The set of vertices of all the subintervals of a subdivision $\{J_1, J_2, \ldots, J_m\}$ will be the partition determined by the subdivision. A partition X will be called regular if the set of vertices of the subintervals in the subdivision determined by X is again X. A partition X_0 will be said to be regularized to X if X consists of the vertices of the subintervals in the subdivision determined by the partition X_0.

For a function $f\colon I \to B$ where B is an ordered Banach space, $I = [\alpha, \beta] \subset R^l$, $\alpha = (\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(l)})$, $\beta = (\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(l)})$ and $\alpha^{(i)} < \beta^{(i)}$ for $i = 1, 2, \ldots, l$ consider the following definitions of the variation of f on I.

Definition 3.1. Let $X = \{x_1, x_2, \ldots, x_n\}$ be called an allowable partition of I if X is such that $x_{k+1} - x_k \in Q_j$ for $k = 1, 2, \ldots, n-1$, Q_j is a fixed positive cone and x_1 is the smallest vertex of I and x_n is the largest vertex of I as determined by Q_j. The variation of f on I, denoted by $V_1(f; I)$, is given by

$$V_1(f; I) = \sup_X \sum_{k=1}^{n-1} \|f(x_k) - f(x_{k+1})\|$$

where the supremum is over all allowable partitions X ordered by a positive cone Q_j for $j = 1, 2, \ldots, 2^{l-1}$. The function f will be said to be of bounded variation, denoted $f \in BV_1$, if $V_1(f; I) < +\infty$.

An interval $I = [\alpha, \beta] \subset R^l$ with the operation $*$ given by

$$x * y = (\max\{x^{(1)}, y^{(1)}\}, \max\{x^{(2)}, y^{(2)}\}, \ldots, \max\{x^{(l)}, y^{(l)}\})$$

forms an idempotent semigroup with identity α. For $x \in I$, the conventions $x^1 = x$ and $x^0 = \alpha$ will be assumed. Thus a special case of the definition of variation given in [9] is given in the following:

Definition 3.2 (Newman). Let T_n denote the Boolean algebra of all n-tuples of zeros and ones. Let $\sigma, \tau \in T_n$ be such that $\tau \geq \sigma$ if $\tau(i) \geq \sigma(i)$ for $i = 1, 2, \ldots, n$ and

$$\mu(\sigma, \tau) = \begin{cases} (-1)^{|\tau| - |\sigma|}, & \tau \geq \sigma, \\ 0, & \text{otherwise,} \end{cases}$$

where $|\sigma|$ denotes the number of ones in the n-tuple σ. Then for $X = \{x_1, x_2, \ldots, x_n\}$ any partition of I with $x_1 = \alpha$ and $x_n = \beta$ define

$$L(X, \sigma)f = \sum_{\tau \in T_n} \mu(\sigma, \tau) f\left(\prod_{i=1}^{n} x_i^{\tau(i)}\right)$$

where Π is the product under the operation $*$ making I an idempotent semigroup. Then the variation of f on I, denoted by $V_2(f; I)$, is given by

$$V_2(f; I) = \sup_X \sum_{\sigma \in T_n} \|L(X, \sigma)f\|$$

where the supremum is taken over all finite partitions X of I. The function f will be said to be of bounded variation, denoted $f \in BV_2$, if $V_2(f; I) < +\infty$.

Definition 3.3 (Arzela). Let $X = \{x_1, x_2, \ldots, x_n\}$ be called a permissible partition of I if X is such that $x_1 = \alpha$, $x_n = \beta$ and $x_{k+1} - x_k \in Q_1$ for $k = 1, 2, \ldots, n-1$. Then the variation of f on I, denoted by $V_3(f; I)$, is given by

$$V_3(f; I) = \sup_X \sum_{k=1}^{n-1} \|f(x_k) - f(x_{k+1})\|$$

where the supremum is taken over all permissible partitions X of I. The function f will be said to be of bounded variation, denoted $f \in BV_3$, if $V_3(f; I) < +\infty$.

Definition 3.4 (Lebesgue). Let $J = [a, b]$ be a subinterval of I and

$$\begin{aligned} \Delta_r f(x^{(1)}, x^{(2)}, \ldots, x^{(r-1)}, x^{(r+1)}, \ldots, x^{(l)}) \\ = f(x^{(1)}, \ldots, x^{(r-1)}, b^{(r)}, x^{(r+1)}, \ldots, x^{(l)}) \\ - f(x^{(1)}, \ldots, x^{(r-1)}, a^{(r)}, x^{(r+1)}, \ldots, x^{(l)}) \end{aligned}$$

for $r = 1, 2, \ldots, l$ and $\Delta_l J = \Delta_1 \Delta_2 \cdots \Delta_l f$. The variation of f on I, denoted $V_4(f; I)$, is given by

$$V_4(f; I) = \sup_Z \sum_{J \in Z} \|\Delta_l J\|$$

where the supremum is taken over all subdivisions Z of I. The function f will be said to be of bounded variation, denoted $f \in BV_4$, if $V_4(f; I) < +\infty$.

If $X = \{x_1, x_2, \ldots, x_n\}$ is a permissible partition for the calculation of $V_3(f; I)$ it is also an allowable partition for $V_1(f; I)$ so the calculation of $V_1(f; I)$ is the supremum over a larger set than is the calculation of $V_3(f; I)$. Thus $f \in BV_1$ implies that $f \in BV_3$. The converse of this statement does not hold however. Consider the function $f: I \to R$, $I = [0, 1] \subset R^2$ defined by

$$f(x, y) = \begin{cases} \sin(1/x) & \text{if } x + y = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Since $\sin(1/x)$ is not of bounded variation on $[0, 1] \subset R$, $f(x, y)$ is not of bounded variation by Definition 1. However any partition ordered by Q_1 contains at most one point where $f(x, y)$ is nonzero so that $V_3(f; I) \leq 2$.

To determine the relationship between $V_2(f; I)$ and $V_4(f; I)$ some preliminary results are necessary. In the calculation of $V_2(f; I)$ the elements of T_1 will be denoted $\sigma_1^{(1)} = (1)$ and $\sigma_1^{(2)} = (0)$. Assume that the elements of T_{n-1} have been numbered, then those of T_n will be given by $\sigma_n^{(1)} = (\sigma_{n-1}^{(1)}, 1)$, $\sigma_n^{(2)} = (\sigma_{n-1}^{(1)}, 0), \ldots, \sigma_n^{(k)} = (\sigma_{n-1}^{(m)}, 1)$ for $k = 2^m - 1$ and $\sigma_n^{(k)} = (\sigma_{n-1}^{(m)}, 0)$ for $k = 2m$ where $k = 1, 2, \ldots, 2^n$. The ith coordinate of $\sigma_n^{(k)}$ will be denoted $\sigma_n^{(k)}(i)$. The subscript n will be deleted from $\sigma_n^{(k)}$ when it is clear that $\sigma_n^{(k)} \in T_n$.

Note 1. Let $f: I \to B$, $I = [\alpha, \beta] \subset R^l$ and $X = \{x_1, x_2, \ldots, x_{2^l}\}$ be the vertices of I with $x_1 = \alpha$, $x_{2^l} = \beta$ and $x_k = (y^{(1)}, y^{(2)}, \ldots, y^{(l)})$ where

$$y^{(i)} = \begin{cases} \alpha^{(i)} & \text{if } t_i = 0, \\ \beta^{(i)} & \text{if } t_i = 1 \end{cases}$$

for $i = 1, 2, \ldots, l$ and $k = 1 + t_1 + 2t_2 + \cdots + 2^{(l-1)} t_l$. Then $L(X, \sigma^{2(2^l-1)})f = \Delta_l I$ where $\Delta_l I$ is as in Definition 5.4.

Proof. Let

$$\text{(1)} \qquad \Gamma_m = \left\{ \sigma^{(k)}: \sigma^{(k)} \in T_{2^l}, \left(\prod_{i=1}^{2^l} x_i^{\sigma^{(k)}(i)} \right) = x_m \right\}$$

for $m = 1, 2, \ldots, 2^l$, let $\sigma^{(p_m)} \in \Gamma_m$ be such that $\sigma^{(k)} \leq \sigma^{(p_m)}$ for all $\sigma^{(k)} \in \Gamma_m$, and let $\eta = \sigma^{2(2^l-1)}$. Since $\sigma^{(k)}(1) = 0$ for $k \geq 2^{(2^l-1)}$ and $\eta(1) = 1$, we have that $\mu(\eta, \sigma^{(k)}) = 0$ for $k > 2^{(2^l-1)}$.

$$L(X, \sigma^{2(2^l-1)})f = \sum_{\sigma \in T_{2^l}} \mu(\sigma^{2(2^l-1)}\sigma) f\left(\prod_{i=1}^{2^l} x_i^{\sigma(i)}\right) = \sum_{\sigma \in \Gamma_1} \mu(\eta, \sigma) f\left(\prod_{i=1}^{2^l} x_i^{\sigma(i)}\right)$$
$$+ \sum_{\sigma \in \Gamma_2} \mu(\eta, \sigma) f\left(\prod_{i=1}^{2^l} x_i^{\sigma(i)}\right) + \cdots + \sum_{\sigma \in \Gamma_{2^l}} \mu(\eta, \sigma) f\left(\prod_{i=1}^{2^l} x_i^{\sigma(i)}\right).$$

Hence it will suffice to show that

$$\sum_{\sigma \in \Gamma_m} \mu(\eta, \sigma) f\left(\prod_{i=1}^{2^l} x_i^{\sigma(i)}\right) = (-1)^{d_m} f(x_m) \tag{2}$$

where d_m is the number of $t_i = 1$ for $m = 1 + t_1 + 2t_2 + \cdots + 2^{(l-1)}t_l$.

For $d_m = 0, 1, 2, \ldots, l$ and $n = 0, 1, \ldots, 2^l - 1$ let $N(d_m, n)$ be the number of distinct elements $\sigma^{(p)} \in \Gamma_m$ such that $\mu(\eta, \sigma^{(p)}) \neq 0$ and $\sigma^{(p)}(i) = 1$ for n values of $i \neq 1$. The number of $i \neq 1$ such that $\sigma^{(p_m)}(i) = 1$ is $2^{d_m} - 1$ and $\sum_{n=0}^{2^l-1} N(d_m, n)$ is the number of elements $\sigma \in \Gamma_m$ such that $\mu(\eta, \sigma) \neq 0$ for $d_m = 0, 1, 2, \ldots, l$. Thus $N(0, 0) = 1$ and

$$N(d_m, n) = \binom{2^{d_m} - 1}{n} - \binom{d_m}{1} N(d_m - 1, n) - \binom{d_m}{2} N(d_m - 2, n)$$
$$- \cdots - \binom{d_m}{d_m - 1} N(1, n) - \binom{d_m}{d_m} N(0, n)$$

for $n \leq 2^{d_m} - 1$ and $N(d_m, n) = 0$ for $n > 2^{d_m} - 1$. Now suppose $\sigma^{(k)} \geq \eta$ and contains $n + 1$ ones so that $\mu(\eta, \sigma^{(k)}) = (-1)^n$. Hence the note will hold if

$$\sum_{j=0}^{2^{l-1}-1} N(d_m, 2j) - \sum_{j=1}^{2^{l-1}} N(d_m, 2j - 1) = (-1)^{d_m} \tag{3}$$

for $d_m = 0, 1, 2, \ldots, l$ since then (2) is true.

For $d_m = 0$ it holds. Assume that (3) holds for $d_m \leq r - 1$. Then for $d_m = r$,

$$\sum_{j=0}^{2^{(2^l-1)}-1} N(r, 2j) - \sum_{j=1}^{2^{l-1}} N(r, 2j-1)$$

$$= \sum_{j=0}^{2^{l-1}-1}\left[\binom{2^r-1}{2j} - \binom{r}{1}N(r-1, 2j) - \binom{r}{2}N(r-2, 2j) - \cdots - \binom{r}{r}N(0, 2j)\right]$$

$$- \sum_{j=1}^{2^{l-1}}\left[\binom{2^r-1}{2j-1} - \binom{r}{1}N(r-1, 2j-1) - \binom{r}{2}N(r-2, 2j-1)\right.$$

$$\left. - \cdots - \binom{r}{r}N(0, 2j-1)\right]$$

$$= -\binom{r}{1}(-1)^{r-1} - \binom{r}{2}(-1)^{r-2} - \cdots - \binom{r}{r-1}(-1)^1 - 1$$

where we define

$$\binom{2^r-1}{h} = 0 \quad \text{if } h > 2^r - 1.$$

Thus we have that

$$-r(-1)^{r-1} - \cdots - \binom{r}{r-1}(-1)^1 - 1 = -[(-1)+1]^r + (-1)^r + 1 - 1 = (-1)^r$$

so the desired result holds.

Note 2. Let $f\colon I \to B$, $I = [\alpha, \beta] \subset R^l$ and let $X = \{x_1, x_2, \ldots, x_{2^l}\}$ be the vertices of I numbered as in Note 1. For Γ_m as in (1), $\sigma^{(p_m)} \in \Gamma_m$; $\sigma^{(p_m)} \geq \sigma^{(k)}$ for all $\sigma^{(k)} \in \Gamma_m$ we have that $L(X, \sigma^{(p_m)})f = \Delta_f I_m$ where I_m is the interval whose vertices are $\{x_i\colon x_i - x_m \in Q_1\}$.

Proof. By definition

$$L(X, \sigma^{p_m})f = \sum_{\sigma^{(k)} \in T_{2^l}} \mu(\sigma^{(p_m)}, \sigma^{(k)}) f\left(\prod_{i=1}^{2^l} x_i^{\sigma^{(k)}(i)}\right).$$

We need only consider those $\sigma^{(k)} \geq \sigma^{(p_m)}$ since $\mu(\sigma^{(p_m)}, \sigma^{(k)}) = 0$ otherwise. The set $\{x_r\colon \prod_{i=1}^{2^l} x_i^{\sigma^{(k)}(i)} = x_r\}$ for some $\sigma^{(k)} \geq \sigma^{(p_m)} = \{x_r\colon x_r - x_m \in Q_1\}$. Now let

$$\Delta_r = \left\{\sigma^{(k)}\colon \prod_{i=1}^{2^l} x_i^{\sigma^{(k)}(i)} = x_r,\quad \sigma^{(k)} \in T_{2^l} \text{ and } \sigma^{(k)} \geq \sigma^{(p_m)}\right\}.$$

By a counting argument similar to that in the proof of Note 1 we have that

$$\sum_{\sigma^{(k)}\in\Delta_r} \mu(\sigma^{(p_m)}, \sigma^{(k)})f\left(\prod_{i=1}^{2l} x_i^{\sigma^{(k)}(i)}\right) = (-1)^{d_r - d_m} f(x_r)$$

where d_r and d_m are the number of i's such that $x^{(i)} = \beta^{(i)}$ in x_r and x_m respectively. Thus $L(X, \sigma^{(p_m)})f = \Delta_f I_m$.

Note 3. Let $f: I \to B$, $I = [\alpha, \beta] \subset R^l$ and $X = \{x_1, x_2, \ldots, x_{2l}\}$ with X the vertices of I numbered as in Note 1. For Γ_m as in (1) and $\sigma^{(p_m)} \in \Gamma_m$, $\sigma^{(p_m)} \geq \sigma^{(k)}$ for all $\sigma^{(k)} \in \Gamma_m$ we have that $L(X, \sigma^{(k)})f = 0$ for $k \neq p_m$.

Proof. Suppose $\sigma^{(j)} \in \Gamma_m$, $\sigma^{(j)} < \sigma^{(p_m)}$.

$$L(X, \sigma^{(j)})f = \sum_{\sigma^{(k)}\in T_{2l}} \mu(\sigma^{(j)}, \sigma^{(k)})f\left(\prod_{i=1}^{2l} x_i^{\sigma^{(k)}(i)}\right)$$

$$= \sum_{\sigma\in\Gamma_m} \mu(\sigma^{(j)}, \sigma)f\left(\prod_{i=1}^{2l} x_i^{\sigma(i)}\right) + \sum_{\sigma\in\Gamma_{m(1)}} \mu(\sigma^{(j)}, \sigma)f\left(\prod_{i=1}^{2l} x_i^{\sigma(i)}\right)$$

$$+ \cdots + \sum_{\sigma\in\Gamma_{m(s)}} \mu(\sigma^{(j)}, \sigma)f\left(\prod_{i=1}^{2l} x_i^{\sigma(i)}\right)$$

such that $x_{m_i} - x_m \in Q_1$ for $i = 1, 2, \ldots, s$. Now $\Sigma_{\sigma\in\ m}\ \mu(\sigma^{(j)}, \sigma)f(\Pi_{i=1}^{2l} x_i^{\sigma(i)})$ contains 2^r nonzero summands where r is the number of i's such that $\sigma^{(p_m)}(i) - \sigma^{(j)}(i) = 1$ and 2^{r-1} of them are such that $\mu(\sigma^{(j)}, \sigma) = +1$ and 2^{r-1} are such that $\mu(\sigma^{(j)}, \sigma) = -1$. Hence

$$\sum_{\sigma\in\Gamma_m} \mu(\sigma^{(j)}, \sigma)f\left(\prod_{i=1}^{2l} x_i^{\sigma(i)}\right) = 0.$$

Also $\Sigma_{\sigma\in\Gamma_{m(k)}}\ \mu(\sigma^{(j)}, \sigma)f(\Pi_{i=1}^{2l} x_i^{\sigma(i)})$ contains 2^s summands where s is the number of i's such that $\sigma^{(p_{m(k)})}(i) - \sigma^{(j)}(i) = 1$ and again 2^{s-1} of them are such that $\mu(\sigma^{(j)}, \sigma) = +1$ and 2^{s-1} are such that $\mu(\sigma^{(j)}, \sigma) = -1$ so that

$$\sum_{\sigma\in\Gamma_{m(i)}} \mu(\sigma^{(j)}, \sigma)f\left(\prod_{i=1}^{r} x_i^{\sigma(i)}\right) = 0$$

and the conclusion holds.

Lemma 3.5. *Let $f: I \to B$, $I = [\alpha, \beta] \subset R^l$ and $X = \{x_1, x_2, \ldots, x_{2l}\}$ be the vertices of I with $x_1 = \alpha$ and $x_{2l} = \beta$. Then*

$$\sum_{\sigma\in T_{2l}} \|L(X, \sigma)f\| = \|\Delta_f I\| + \sum_{J\in\Omega} \|\Delta_f J\|.$$

Proof. By definition

$$\sum_{\sigma\epsilon T_{2^l}} \|L(x,\sigma)f\| = \sum_{n=1}^{2^l} \|L(X,\sigma^{(p_m)})f\| + \sum_{\sigma^{(k)}\neq\sigma^{(p_m)}} \|L(X,\sigma^{(k)})f\|.$$

By Note 3 $L(X,\sigma^{(k)})f = 0$ for $k \neq p_m$ for some m so that

$$\sum_{\sigma\epsilon T_{2^l}} \|L(X,\sigma)f\| = \sum_{m=1}^{2^l} \|L(X,\sigma^{(p_m)})f\|.$$

But by Note 1, $L(X,\sigma^{2(2^l-1)})f = \Delta_f I$ and by Note 2, $L(X,\sigma^{(p_m)})f = \Delta_f I_m$ so the lemma holds.

The following lemmas yield the equivalence of $V_2(f; I)$ and $V_4(f; I)$ provided $V_2(f; J) = 0$ for $J \in \Omega$.

Lemma. *Let $f: I \to B$, $I = [\alpha, \beta] \subset R^l$. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a partition of I and $X_0 = X \cup \{y\}$, $y \notin X$, be a second partition of I. Then*

$$\sum_{\sigma\epsilon T_n} \|L(X,\sigma)f\| \leq \sum_{\sigma\epsilon T_{n+1}} \|L(X_0,\sigma)f\|.$$

This lemma follows since $L(X,\sigma^{(k)})f = L(X_0,\tau_1)f + L(X_0,\tau_2)f$ where

$$\tau_1(i) = \begin{cases} \sigma^{(k)}(i) & \text{for } i < m, \\ 0 & \text{for } i = m, \\ \sigma^{(k)}(i-1) & \text{for } i > m, \end{cases}$$

and

$$\tau_2(i) = \begin{cases} \tau_1(i) & \text{for } i \neq m, \\ 1 & \text{for } i = m. \end{cases}$$

Lemma 3.6. *Let $f: I \to B$, $I = [\alpha, \beta] \subset R^l$ and $X = \{x_1, x_2, \ldots, x_r\}$ be a regular partition of I with $\{I_1, I_2, \ldots, I_s\}$ the corresponding regular subdivision of I in the calculation of $V_4(f; I)$. Then*

$$\sum_{\sigma\epsilon T_r} \|L(X,\sigma)f\| = \sum_{m=1}^{s} \|\Delta_f I_m\|$$

assuming the variation, $V_2(f; J) = 0$ for $J \in \Omega$.

The proof follows in a manner similar to that of Note 1.

Lemma 3.7. *Let $f: I \to B$, $I = [\alpha, \beta] \subset R^l$ and $V_2(f; J) = 0$ for $J \in \Omega$. Then $f \in BV_2$ if and only if $f \in BV_4$.*

This follows from lemmas above.

It remains to establish a relationship between V_1 and either V_2 or V_4.

Note 4. Let $f\colon I \to B$, $I = [\alpha, \beta] \subset R^l$, $f \in BV_2$ and $V_2(f; I) = 0$ for $J \in \Omega$. Then $f \in BV_1$.

Proof. From Lemma 3.7, BV_2 and BV_4 are equivalent when $V_2(f; J) = 0$ for $J \in \Omega$. The calculation of V_4 is symmetrical and the calculations of V_2 and V_1 agree when the partition is ordered by Q_1 so the result follows.

The converse of Note 4 does not hold. Consider the function $f\colon I \to R$, $I = [0, 1] \subset R^2$ defined by

$$f(x, y) = \begin{cases} (-1)^{n+k}\left(\dfrac{1}{2^n}\right) & \text{if } x = \dfrac{k}{2^n},\ y = \dfrac{1}{2^n};\ x = \dfrac{k}{2^n},\ y = 1 - \dfrac{1}{2^n} \\ & \quad \text{for } k = 1, 2, \ldots, 2^n - 1, \\ (-1)^{n+j}\left(\dfrac{1}{2^n}\right) & \text{if } x = \dfrac{1}{2^n},\ y = \dfrac{j}{2^n};\ x = 1 - \dfrac{1}{2^n},\ y = \dfrac{j}{2^n} \\ & \quad \text{for } j = 1, 2, \ldots, 2^n - 1, \\ 0 & \text{otherwise.} \end{cases}$$

For N any positive integer, the regular partition $P = \{(k/2^N, 1/2^N)\colon k, j = 0, 1, 2, \ldots, 2^N\}$ we have that

$$\begin{aligned} V_2(f; I) &\geq \sum_{j=0}^{2^N} \sum_{k=0}^{2^N} \left| f\left(\frac{k}{2^N}, \frac{j}{2^N}\right) - f\left(\frac{k+1}{2^N}, \frac{j}{2^N}\right) - f\left(\frac{k}{2^N}, \frac{j+1}{2^N}\right) + f\left(\frac{k+1}{2^N}, \frac{i+1}{2^N}\right) \right| \\ &\geq \frac{1}{2} + 2 + 3(N-2) \end{aligned}$$

since

$$f\left(\frac{k}{2^N}, \frac{j}{2^N}\right) f\left(\frac{k+1}{2^N}, \frac{j+1}{2^N}\right) \geq 0, \qquad f\left(\frac{k+1}{2^N}, \frac{j}{2^N}\right) f\left(\frac{k}{2^N}, \frac{j+1}{2^N}\right) \geq 0$$

and

$$f\left(\frac{k}{2^N}, \frac{j}{2^N}\right) f\left(\frac{k+1}{2^N}, \frac{j}{2^N}\right) \leq 0.$$

Hence $V_2(f; I) > N$ for $N \geq 2$.

On the other hand for any partition $P \in \mathcal{P}$, $\mathcal{P}$ the set of all finite nonempty partitions linearly ordered by either Q_1 or Q_2, we have $V_1(f; I) \leq [2(2^N) - 1](1/2^N) \leq 2$ since $\mathcal{P}$ contains at most $2(2^n) - 1$ points of the form $(k/2^n, j/2^n)$ with a total number less than or equal to $2(2^N) - 1$ so that $f \in BV_1$.

The example showing that $f \in BV_3$ does not imply $f \in BV_1$ also shows

then that $f \in BV_3$ does not imply $f \in BV_2$ or $f \in BV_4$.

4. **Primary results.** We now address the question of giving conditions which when satisfied by the sequence $\{y_k\}$, $y_k \in \Pi_{s \in S} E_s$, where S is a nonempty set in R^l and each E_s is an ordered vector space, guarantee the existence of a subsequence of $\{y_k\}$ which converges pointwise on S. With this in mind we make the following definition.

Definition 4.1. Let S be a nonempty subset of R^l and f be a function, $f \in \Pi_{s \in S} E_s$ where each E_s is an ordered vector space with positive cone K_s. Consider the set $\mathcal{P}$ of all finite nonempty partitions $P = \{x_1, x_2, \ldots, x_n\}$ of S where $n \geq 1$, $x_i \in S$ for $i = 1, 2, \ldots, n$ and P is linearly ordered by some positive cone Q_j for $j = 1, 2, \ldots, 2^{(l-1)}$. If $f(s) \neq \theta_s$ for some $s \in S$ we say that (f, P) is a proper pair if $(-1)^i f(x_i) > \theta_{x_i}$ for $i = 1, 2, \ldots, n$ or else $(-1)^i f(x_i) < \theta_{x_i}$ for $i = 1, 2, \ldots, n$. If $f(s) = \theta_s$ for all $s \in S$ we say that (f, P) is a proper pair if P contains exactly one point.

We will say that the sequence is an eventually comparable sequence if there exists a positive integer $M(t)$ such that $y_k(t)$ and $y_j(t)$ are comparable for $k, j \geq M(t)$.

Theorem 4.2. *Let S be a nonempty subset of R^l and $\{y_k\}$ be a sequence of functions, $y_k \in \Pi_{s \in S} E_s$, where each E_s is an ordered vector space with positive cone K_s. For each $t \in S$ assume that $\{y_k(t)\}$ is an eventually comparable sequence and that there is a number $\delta(t) \geq 0$ and positive integers $N(t)$ and $H(t)$ such that for each $k, j \geq H(t)$ the partitions $P \in \mathcal{P}$ such that $(y_k - y_j, P)$ is a proper pair, each contain at most $N(t)$ points in the sphere $D(t, \delta(t))$. Then $\{y_k\}$ contains a subsequence $\{h_k\}$ such that $\{h_k(t)\}$ is eventually monotone for each $t \in S$.*

Proof. If $y_k(t)$ and $y_j(t)$ are comparable for all $k, j \geq M(t)$ and $M(t)$ is the smallest integer having this property we let $A_i = \{t: t \in S, M(t) = i\}$ for $i = 1, 2, \ldots$. For any $z \in A_i$ we have $y_k(t)$ and $y_j(t)$ comparable for $k, j \geq i$. We will prove the theorem assuming that $y_k(t)$ and $y_j(t)$ are comparable for all $t \in S$ and then a standard disagonalization argument where S is replaced by $A_1, A_2, \cdots$ yields the desired result.

We note that we may assume S is bounded because if the theorem is true for bounded sets a standard diagonalization argument with $S \cap D(0, 1)$, $S \cap D(0, 2), \cdots$ yields the result for unbounded sets. Also, we may assume S is a closed interval because if the theorem is true for closed intervals $I = [\alpha, \beta]$, then we may choose I to be a closed interval containing the bounded set S and define a sequence of functions $\{Z_k\}$, $Z_k \in \Pi_{s \in I} F_s$ where $F_s = E_s$ for $s \in S$ and $F_s = R$ (with the usual order) for $s \in I$, $s \notin S$, by

$$Z_k(t) = \begin{cases} y_k(t) & \text{for } t \in S, \\ 0 & \text{for } t \notin S. \end{cases}$$

Then the sequence $\{Z_k\}$ satisfies the hypotheses of the theorem on I since for $(Z_k - Z_j, P)$ a proper pair either $Z_k(t) - Z_j(t) = 0$ for all $t \in I$ or else P contains no points in $I - S$. Also, since we may assume S is a compact interval there exists a finite number of $t_i \in S$, $i = 1, 2, \ldots, n$, such that $D(t_i, \delta(t_i))$, $i = 1, 2, \ldots, n$, cover S and by choosing $H = \max_{1 \leq i \leq n} \{H(t_i)\}$ and $N = \sum_{i=1}^n N(t_i)$ we have integers N and H such that for $k, j \geq H$ the partitions $P \in \mathcal{P}$ such that $(y_k - y_j, P)$ is a proper pair, each contain at most N points in S. Hence without loss of generality we may assume $H = 1$.

The theorem will now be proved by induction on l for S a compact interval in R^l.

Let $l = 1$, S be a compact interval and $\{y_k\}$ be a sequence of functions $y_k \in \prod_{s \in S} E_s$ where E_s is an ordered vector space with positive cone K_s. For each k, j and $t \in S$ assume that $y_k(t)$ is comparable with $y_j(t)$. Let N be a positive integer such that for each k, j the partitions $P \in \mathcal{P}$ with $(y_k - y_j, P)$ a proper pair each contain at most N points in S. The case for $l = 1$ will now be proved by induction on N. If $N = 1$ then for each k, j either $y_k(t) - y_j(t) \in K_t$ holds for all $t \in S$ or else $y_j(t) - y_k(t) \in K_t$ holds for all $t \in S$. Thus by the mapping g, defined in the proof of Corollary 2.2, it is possible to pick a subsequence of $\{y_k\}$ that is monotone on S. Now assume the theorem holds for $N = 1, 2, \ldots, K$. We will then show this implies it is correct for $N = K + 1$.

If there are only finitely many functions in $\{y_k\}$ which are distinct on S then infinitely many are identical on S and using that as the subsequence we are done. Thus we may assume there are infinitely many functions in $\{y_k\}$ which are distinct on S and, by picking a subsequence if necessary, we may assume all the y_k are distinct on S.

Let I, J be subintervals of S with $I \cup J = S$, $I \cap J = \emptyset$ and let I and J be of the same length. We will show that there is a subsequence of $\{y_k\}$ that is eventually monotone for each $t \in I$ or else is eventually monotone for each $t \in J$. Now let $\Gamma = \{y_k\}$, $u = K + 1$ and $r = 2$ with $C_1 = \{\{y_k, y_j\}$: $k \neq j$, proper pairs $(y_k - y_j, P)$ are such that P has at most one point $x \in I\}$. Now for $n = 2, 3, \ldots, K + 1$ let $C_n = \{\{y_k, y_j\}$: $k \neq j$, proper pairs $(y_k - y_j, P)$ are such that P has at most n points $x_1, x_2, \ldots, x_n \in I$ and $\{y_k, y_j\} \notin C_{n-p}$, $p = 1, 2, \ldots, n-1\}$. By Theorem 2.1 there is an infinite subclass Δ of Γ such that all pairs of elements of Δ belong to the same C_m for some $m = 1, 2, \ldots, K + 1$. If $m = 1, 2, \ldots, K$ then the induction hypothesis holds on

I and there is a subsequence $\{h_j\}$ of $\{y_k\}$ that is eventually monotone at each point of I. Note that if $\{y_k, y_j\} \in C_m$ then proper pairs $(y_k - y_j, P)$ are such that P has at most $K + 2 - m$ points $x_1, x_2, \ldots, x_{K+2-m} \in J$ so that if $m = 2, 3, \ldots, K$ a similar argument shows that the induction hypothesis holds on J so there is a subsequence $\{g_i\}$ of $\{h_j\}$ that is eventually monotone at each point of $I \cup J = S$ and the case $l = 1$ holds. If $m = K + 1$ then the proper pairs $(y_k - y_j, P)$ are such that P has at most one point $x \in J$ and there is a subsequence $\{h_j\}$ of $\{y_k\}$ that is eventually monotone at each point of J. In any case either the case $l = 1$ holds or there is a subsequence that is eventually monotone at each point of I or J. Let S_1 denote the interval I or J on which $\{h_j\}$ is not known to be eventually monotone at each point.

We now repeat the entire process described in the preceding paragragh on S_1 obtaining S_2, then on S_2 obtaining S_3, etc. If the theorem does not hold at any step in the induction, use the standard diagonalization argument to obtain a sequence $\{h_j\}$. Let x_0 be the unique limit point of the set of midpoints of the intervals S_n. Choose $\{g_i\}$ to be a subsequence of $\{h_j\}$ such that $\{g_i(x_0)\}$ is monotone so $\{g_i\}$ now has the property that it is eventually monotone at each point of S and the case for $l = 1$ holds.

Assume the theorem is true for $l = L - 1$ so it remains to show that it holds for $l = L$.

Let $S = [\alpha, \beta]$ be a compact interval in R^L and $\{y_k\}$ be a sequence of functions, $y_k \in \Pi_{s \in S} E_s$, where each E_s is an ordered vector space with positive cone K_s. For each k, j and $t \in S$ assume that $y_k(t)$ is comparable with $y_j(t)$. Let N be a positive integer such that for each k, j the partitions $P \in \mathcal{P}$ with $(y_k - y_j, P)$ a proper pair each contain at most N points in S. Now the case for $l = L$ will be proved by induction on N. If $N = 1$ then for each k, j either $y_k(t) - y_j(t) \in K_t$ holds for all $t \in S$ or else $y_j(t) - y_k(t) \in K_t$ holds for all $t \in S$. Thus it is possible to pick a subsequence as in the proof of Corollary 2.2 which is a monotone sequence on S. We now assume the theorem holds for $N = 1, 2, \ldots, K$ and will show this implies it is correct for $N = K + 1$.

If there are only finitely many functions in $\{y_k\}$ which are distinct on S then infinitely many are identical on S and we are done. Thus we may assume there are infinitely many functions in $\{y_k\}$ that are distinct on S and, by picking a subsequence if necessary, we may assume all the y_k's are distinct. Let I, J be subintervals of S with $I \cup J = S$, $I \cap J = \emptyset$, $I = [\alpha, \gamma_1] - X_0$, $J = [\gamma_2, \beta]$ where $\gamma_1 = (\frac{1}{2}(\alpha^{(1)} + \beta^{(1)}), \beta^{(2)}, \beta^{(3)}, \ldots, \beta^{(L)})$, $\gamma_2 = (\frac{1}{2}(\alpha^{(1)} + \beta^{(1)}), \alpha^{(2)}, \alpha^{(3)}, \ldots, \alpha^{(L)})$ and $X_0 = \{(\frac{1}{2}(\alpha^{(1)} + \beta^{(1)}), x^{(2)}, \ldots, x^{(L)}) \colon \alpha^{(i)} \leq x^{(i)}$

$\leq \beta^{(i)}$ for $i = 2, 3, \ldots, L\}$. We will show by the same argument as for $l = 1$ that there is a subsequence of $\{y_k\}$ that is eventually monotone for each $t \in I$ or else is eventually monotone for each $t \in J$. Since all $y_k(\beta)$ are comparable there is a subsequence of $\{y_k\}$, denoted again by $\{y_k\}$, such that $\{y_k(\beta)\}$ is monotone. Now using Theorem 2.1 with $\Gamma = \{y_k\}$, $u = K + 1$ and $r = 2$ where $C_1 = \{\{y_k, y_j\}: k \neq j$, proper pairs $(y_k - y_j, P)$ are such that P contains at most one point $x \in I\}$ and $C_n = \{\{y_k, y_j\}: k \neq j$, proper pairs $(y_k - y_j, P)$ are such that P has at most n points $x_1, x_2, \ldots, x_n \in I$ and $\{y_k, y_j\} \notin C_{n-p}$, $p = 1, 2, \ldots, n-1\}$ for $n = 2, \ldots, K+1$ there is an infinite subclass Δ of Γ such that all pairs of elements of Δ belong to the same C_m for some $m = 1, 2, \ldots, K+1$. If $m = 1, 2, \ldots, K$ then the induction hypothesis holds on I and there is a subsequence $\{h_j\}$ of $\{y_k\}$ such that $\{h_j\}$ is monotone at each point of I. Again if $\{y_k, y_j\} \in C_m$ then the proper pairs $(y_k - y_j, P)$ are such that P has at most $K + 2 - m$ points $x_1, x_2, \ldots, x_{K+2-m} \in J$. So if $m = 2, 3, \ldots, K$ then the induction hypothesis holds on J so there is a subsequence $\{g_i\}$ of $\{h_j\}$ that is monotone at each point of $I \cup J = S$ and the theorem holds. If $m = K + 1$ then the proper pairs $(y_k - y_j, P)$ are such that P has at most one point $x \in J$ and there is a subsequence $\{h_j\}$ of $\{y_k\}$ that is monotone at each point of J. In any case either the theorem holds or there is a subsequence that is monotone at each point of I or is monotone at each point of J. Let S_1 denote the interval I or J on which $\{h_j\}$ is not known to be eventually monotone.

Now repeat the entire process described in the preceding paragraph on S_1 to obtain S_2 so that we obtain a sequence that is eventually monotone on $(S - S_2)$. Continuing, we repeat the entire process on S_2 to obtain S_3, on S_3 to obtain S_4, etc. If the theorem does not hold at any step in the induction, then by a standard diagonalization argument we obtain a sequence $\{h_j\}$ that is eventually monotone on $S_0 = (S - \bigcap_{j=1}^{+\infty} S_i)$. But $\bigcap_{i=1}^{+\infty} S_i$ is an $L - 1$ dimensional interval so by the induction hypothesis on the dimension of the domain it is possible to choose $\{g_i\}$, a subsequence of $\{h_j\}$, such that $\{g_i(x)\}$ is eventually monotone for $x \in \bigcap_{i=1}^{+\infty} S_i$. Now $\{g_i\}$ has the property that it is eventually monotone at each point of $S \subset R^L$ so the conclusion of the theorem holds.

In the next theorem, by further restricting E_s, we obtain a subsequence $\{h_j\}$ of $\{y_k\}$ such that $\{h_j(s)\}$ is eventually monotone in E_s or else converges in E_s for each $s \in S$. The proof will again be done by induction on the dimension of the domain. The following is the case for $l = 1$.

Theorem 4.3. *Let S be a nonempty subset of real numbers and $\{y_k\}$ be a sequence of functions, $y_k \in \Pi_{s \in S} E_s$, where each E_s is a sequentially com-*

plete ordered locally convex space with positive cone K_s. *For each* $t \in S$ *assume that* $\{y_k(t)\}$ *is an eventually comparable sequence. Assume* E_s *has a nested countable basis of circled sets at* θ_s *denoted by* $\{U_s(n)\}$. *For each* $t \in S$ *and each positive integer* n *assume that there are nonnegative integers* $N(n, t)$, $H(n, t)$ *and a number* $\delta(n, t) > 0$ *such that for all* $k, j \geq H(n, t)$ *if* $(y_k - y_j, P)$ *is a proper pair then* P *contains at most* $N(n, t)$ *points* x *such* $y_k(x) - y_j(x) \notin U_x(n)$ *and* $t - \delta(n, t) < x < t + \delta(n, t)$. *Then* $\{y_k\}$ *contains a subsequence* $\{h_j\}$ *such that for each* $t \in S$, $\{h_j(t)\}$ *is either eventually monotone or else is convergent.*

Proof. See note following Theorem 2.2 in [15].

Now for the case $l > 1$ the following theorem yields the desired conclusion.

Theorem 4.4. *Let* S *be a nonempty subset of* R^l *and* $\{y_k\}$ *be a sequence of functions,* $y_k \in \Pi_{s \in S} E_s$, *where each* E_s *is a sequentially complete ordered locally convex space with positive cone* K_s. *For each* $t \in S$ *assume that* $\{y_k(t)\}$ *is an eventually comparable sequence. Assume* E_s *has a nested countable basis of circled sets of* θ_s *denoted by* $\{U_s(n)\}$. *For each* $t \in S$ *and each positive integer* n *assume that there are nonnegative integers* $N(n, t)$, $H(n, t)$ *and a number* $\delta(n, t) > 0$ *such that for all* $k, j > H(n, t)$ *if* $(y_k - y_j, P)$ *is a proper pair then* P *contains at most* $N(n, t)$ *points* x *such that* $y_k(x) - y_j(x) \notin U_x(n)$ *and* x *is in the sphere* $D(t, \delta(n, t))$. *Then* $\{y_k\}$ *contains a subsequence* $\{h_k\}$ *such that for each* $t \in S$, $\{h_k(t)\}$ *is either eventually monotone or else is convergent.*

Proof. As in the proof of Theorem 3.1 we may assume that $y_k(t)$ and $y_j(t)$ are comparable for all k, j and $t \in S$. Also we will again assume S is a bounded closed interval, $S = [\alpha, \beta]$ and $H(n, t)$ will be replaced by $H(n)$ and $N(n)$ respectively.

The proof now will be by induction on the dimension of the domain. The case for $l = 1$ is Theorem 3.2. Assume the result holds for $l = K - 1$ then we will show that the theorem is true for $l = K$.

Let $\{J_i\}$ be an enumeration of the spheres $D(t, \delta) \subset S$ in R^K with rational centers and rational radii.

Now for J_1 by Corollary 2.2 there is a subsequence of $\{y_k\}$, again denoted $\{y_k\}$, that is monotone on J_1 or else there is a subsequence of $\{y_k\}$ again denoted by $\{y_k\}$ such that for $k \neq j$, $y_k(t) > y_j(t)$ for some $t \in J_1$ and $y_k(r) < y_j(r)$ for some $r \in J_1$. Now repeat the process described in the preceding sentence for $J_2, J_3, \cdots$ and then take the diagonal subsequence, denote it again by $\{y_k\}$. This sequence has the property that on $\{J_i\}$ it is either eventually

monotone or else for every $k \neq j$ sufficiently large, depending on i, there exists $t, r \in J_i$ such that $y_k(t) > y_j(t)$ and $y_k(r) < y_j(r)$.

Now using J_1 and $U_t(1)$ by Corollary 2.3 there exists a subsequence of $\{y_k\}$, again denoted by $\{y_k\}$, such that, for $k \neq j$, $y_k(t) - y_j(t) \in U_t(1)$ for all $t \in J_1$ or else there is a subsequence of $\{y_k\}$, again denoted by $\{y_k\}$, such that, for $k \neq j$ there is a $t \in J_1$ with $y_k(t) - y_j(t) \notin U_t(1)$. Now repeat the process described in the previous sentence using $U_t(2), U_t(3), \cdots$ and then take the diagonal subsequence and denote it again by $\{y_k\}$. This sequence has the property that for J_1 and $U_t(n)$ either for all $k \neq j$ sufficiently large, depending on n, $y_k(t) - y_j(t) \in U_t(n)$ for all $t \in J_1$ or else for all $k \neq j$ sufficiently large depending on n, there is some $t \in J_1$ such that $y_k(t) - y_j(t) \notin U_t(n)$.

We now repeat the entire process described in the preceding paragraph consecutively on the spheres $J_2, J_3, \cdots$ and then take the diagonal subsequence again denoted by $\{y_k\}$. This sequence has the property that for J_i and $U_t(n)$ either for $k \neq j$ sufficiently large depending on i and n, $y_k(t) - y_j(t) \in U_t(n)$ for every $t \in J_i$ or else for every $k \neq j$ sufficiently large, depending on i and n, there exists a $t \in J_i$ such that $y_k(t) - y_j(t) \notin U_t(n)$.

Let $\Gamma_1, \Gamma_2, \ldots, \Gamma_{2K}$ be the bounding edges of S; then Γ_1 is a $K-1$ dimensional interval so by the induction hypothesis there is a subsequence of $\{y_k\}$, denote it again by $\{y_k\}$, such that for $x \in \Gamma_1$, $\{y_k(x)\}$ is either eventually monotone or else is convergent. Then by similarly extracting a subsequence from $\{y_k\}$ on $\Gamma_2, \Gamma_3, \ldots, \Gamma_{2K}$ consecutively we obtain a subsequence that is either eventually monotone or convergent for each point of the bounding edges of S.

Now suppose A_0 is the set of x's in S such that $\{y_k(x)\}$ does not converge and is not eventually monotone. If A_0 is countable then by a standard diagonalization argument there is a subsequence of $\{y_k\}$, denote it again by $\{y_k\}$, that is either eventually monotone or else is convergent for each $x \in S$. So suppose A_0 is uncountable. Let Γ be the set of $K-1$ dimensional hyperplanes of the form $\gamma = \{(x^{(1)}, x^{(2)}, \ldots, x^{(K)}): x^{(i)} = c$ for i fixed, $\alpha^{(i)} \leq c \leq \beta^{(i)}\}$ which have nonempty intersection with A_0. Consider the set Ψ of sets B of the form

$$B = \left\{(\gamma, a_\gamma): \gamma \in \Gamma,\ a_\gamma \in A_0 \cap \gamma,\ a_{\gamma_1} - a_{\gamma_2} \in \bigcup_{i=1}^{2K-1} [Q_i^0 \cup (-Q_i^0)]\right\}.$$

Now order the sets B in Ψ by $B_1 \leq B_2$ if and only if $B_1 \subset B_2$. Then the union of any linearly ordered chain in Ψ is an upper bound of the chain and is again in Ψ so by Zorn's lemma there must be a maximal element B_0 in Ψ.

If B_0 is uncountable and $(\gamma_1, a_{\gamma_1}), (\gamma_2, a_{\gamma_2})$ are in B_0 then $(a_{\gamma_1} -$

$a_{\gamma_2}) \in \bigcup_{i=1}^{2^{K-1}} [Q_i^0 \cup (-Q_i^0)]$. Let $A_1 = \{a_\gamma : (\gamma, a_\gamma) \in B_0\}$. Now for $x \in A_1$ let $\{F_{xi}\}$ be the subsequence of $\{J_i\}$ consisting of the spheres which contain x. There must be a smallest positive integer n_{xi}, such that $y_k(t) - y_j(t) \notin U_t(n_{xi})$ for all $k \neq j$ sufficiently large, depending on i, for some $t \in F_{xi}$ for otherwise $\{y_k\}$ would be Cauchy on F_{xi} and hence would be convergent at x which is contrary to the choice of x. If $\overline{\lim}_{i\to+\infty} n_{xi} = +\infty$ then there is a subsequence $\{n_{xi(\alpha)}\}$ of $\{n_{xi}\}$ such that $\lim_{\alpha\to+\infty} n_{xi(\alpha)} = +\infty$ and by the definition of $n_{xi(\alpha)}$ and the nestedness of $\{U_t(n)\}$ we have that $y_k(t) - y_j(t) \in U_t(n_{xi(\alpha)} - 1)$ for all $k \neq j$ sufficiently large, depending on α, and all $t \in F_{xi(\alpha)}$. This implies that $\{y_k(x)\}$ is Cauchy and hence convergent which is contrary to the choice of x so that $\overline{\lim}_{i\to+\infty} n_{xi} = c_x < +\infty$. Let $c_x < d_x$ be an upper bound for the set $\{n_{xi}\}$.

Since there are uncountably many values of x in A_1 at which $\{y_k(x)\}$ is not convergent nor eventually monotone then there is some fixed positive integer d such that $d_x \leq d$ holds for uncountably many values of x in A_1. Call this set A so that A is an infinite set of distinct elements. Let $\Gamma_0 = A$, $u = 2^{K-1}$ and $r = 2$ with $C_m = \{\{a_{\gamma_k}, a_{\gamma_j}\}: a_{\gamma_k} - a_{\gamma_j} \in ((Q_m^0) \cup (-Q_m^0))\}$ for $m = 1, 2, 3, \ldots, 2^{K-1}$. Then by Theorem 2.1 there is an infinite subclass Δ of Γ_0 such that for $x, y \in \Delta, x - y \in (-Q_m^0 \cup (-Q_m^0))$ for m fixed. Also for $x \in \Delta$ and $k \neq j$ sufficiently large, depending on i, there exists a $t \in F_{xi}$ such that $y_k(t) - y_j(t) \notin U_t(d)$.

Choose $N > N(d)$ and $u(1) \in \Delta \cap S^0$ and $F_{u(1)i(1)} \in \{F_{u(1)i}\}$ such that $(S - F_{u(1)i(1)}) \cap \Delta$ is infinite. Choose $u(2) \in (S - F_{u(1)i(1)}) \cap (\Delta \cap S^0)$ and $F_{u(2)i(2)} \in \{F_{u(2)i}\}$ such that $F_{u(1)i(1)} \cap F_{u(2)i(2)} = \emptyset$ and $(S - (F_{u(1)i(1)} \cup F_{u(2)i(2)})) \cap \Delta$ is infinite. Continuing in this manner we get $\{u(1), u(2), \ldots, u(2N+1)\}$ in $\Delta \cap S^0$ and $\{F_{u(1)i(1)}, F_{u(2)i(2)}, \ldots, F_{u(2N+1)i(2N+1)}\}$ which are mutually disjoint. Now by renaming if necessary we may assume that $u(j+1) - u(j) \in Q_m^0$ for $j = 1, 2, \ldots, 2N$. Let

$$p = \min_{1\leq i\leq 2N} \{|u^{(1)}(i) - u^{(1)}(i+1)|, |u^{(2)}(i) - u^{(2)}(i+1)|, \ldots, |u^{(K)}(i) - u^{(K)}(i+1)|\}$$

so $p > 0$. Choose $G_{u(j)i(j)} \in \{F_{u(j)i}\}$ such that $G_{u(j)i(j)} \subset F_{u(j)i(j)}$ and the radius of $G_{u(j)i(j)}$ is less than $p/2$. Now $\{G_{u(1)i(1)}, G_{u(2)i(2)}, \ldots, G_{u(2N+1)i(2N+1)}\}$ are mutually disjoint and if $x \in G_{u(j)i(j)}$, $y \in G_{u(j+1)i(j+1)}$ then $y - x \in Q_m^0$. Choose $k \neq j$, k, $j > H(d)$, sufficiently large that for each odd positive integer α, $1 \leq \alpha \leq 2N+1$, $y_k(x_\alpha) - y_j(x_\alpha) \notin U_{x_\alpha}(d)$ for some $x_\alpha \in G_{u(\alpha)i(\alpha)}$ and for each positive even integer α, $2 \leq \alpha \leq 2N$, $y_k(t_\alpha) - y_j(t_\alpha)$

$< \theta$ holds for some $t_\alpha \in G_{u(\alpha)i(\alpha)}$ and $y_k(r_\alpha) - y_j(r_\alpha) > \theta$ holds for some $r_\alpha \in G_{u(\alpha)i(\alpha)}$. Now consider the partition $P_0 = \{\beta_1, \beta_2, \dots, \beta_n\}$ where $\beta_\alpha = x_\alpha$ if α is odd; β_α is omitted from P_0 if α is even and either $y_k(x_{(\alpha-1)}) - y_j(x_{(\alpha-1)}) < \theta$ and $y_k(x_{(\alpha+1)}) - y_j(x_{(\alpha+1)}) > \theta$ or the opposite inequalities hold; β_α is taken to be t_α if $y_k(x_{(\alpha-1)}) - y_j(x_{(\alpha-1)}) > \theta$ and $y_k(x_{(\alpha+1)}) - y_j(x_{(\alpha+1)}) > \theta$ and β_α is taken to be r_α if $y_k(x_{(\alpha-1)}) - y_j(x_{(\alpha-1)}) < \theta$ and $y_k(x_{(\alpha+1)}) - y_j(x_{(\alpha+1)}) < \theta$. Then the partition P_0 is such that $(y_k - y_j, P_0)$ is a proper pair and $y_k(x_\alpha) - y_j(x_\alpha) \notin U_{x_\alpha}(d)$ for α odd, $x_\alpha \in P_0$, and there are $N+1$ such points x_α which is contrary to the hypothesis of the theorem. Hence it cannot be the case that B_0 is uncountable.

Let $B_0 = \{(\gamma_1, a_{\gamma_1}), (\gamma_2, a_{\gamma_2}), \cdots\}$ and ρ_i, $i = 1, 2, \dots$ be the countable number of regular $K-1$ dimensional hyperplanes through the points of $a_{\gamma_1}, a_{\gamma_2}, \cdots$. Suppose $\gamma \in \Gamma$ with $\gamma \neq \rho_i$, $i = 1, 2, \cdots$ and $a_{\gamma_i} \notin \gamma$ for $i = 1, 2, \cdots$ then $[\bigcap_{i=1}^{+\infty} (\gamma \cap \tilde{\rho}_i)] \cap A_0 = \emptyset$ since otherwise B_0 would not be maximal. Then $A_0 \subset \bigcup_{i=1}^{+\infty} \rho_i$ so apply the induction hypothesis to the sequence $\{y_k\}$ on $\rho_1, \rho_2, \cdots$ consecutively, take the diagonal subsequence and denote it again by $\{y_k\}$. This sequence has the property that it is either convergent or eventually monotone at every point in S.

5. Corollaries and other results. By restricting the range of the sequence of functions so that eventually monotone sequences in E_s converge in E_s the conclusion of Theorem 4.4 can be changed to read pointwise convergent. The following is an example of such a theorem.

Theorem 5.1. *Let B be a reflexive ordered Banach space with normal positive cone K and S be a nonempty subset of R^l. If $\{y_k\}$ is a sequence of functions that satisfies the hypothesis of Theorem 4.4 and $\{y_k(s)\}$ is a norm bounded set for each $s \in S$ then there is a subsequence of $\{y_k\}$ which converges for every $s \in S$.*

Proof. This follows directly from Theorem 4.4.

Theorem 5.2 (Helly selection theorem). *Let $\{y_k\}$ be a sequence of functions, $y_k: I \to R$, $I = [\alpha, \beta] \subset R^l$. Suppose there is a positive constant K with $\{y_k(x)\}$ such that $|y_k(x)| \leq K$ for $k = 1, 2, \cdots$ and $x \in I$ and $V_1(y_k; I) \leq K$ for $k = 1, 2, \cdots$. Then there is a subsequence $\{h_k\}$ of $\{y_k\}$ that converges pointwise on I.*

Proof. Since $V_1(y_k - y_j; I) \leq V_1(y_k; I) + V_1(y_j; I) \leq 2K$ the hypotheses of Theorem 5.1 are satisfied and the conclusion holds.

Since $f \in BV_4$ and $f \in BV_2$ implies $f \in BV_1$ the above Helly selection theorem holds if $V_1(y_k; I) \leq K$ is changed to $V_4(y_k; I) \leq K$ or $V_2(y_k; I) \leq K$.

In the case where the sequence of functions $\{y_k\}$ is such that $y_k: S \to R$, $S \subset R^l$, we obtain a characterization of those sequences of functions which have pointwise convergent subsequences.

Definition 5.3. Let S be a nonempty subset of R^l and $\{y_k\}$ be a sequence of functions, $y_k: S \to R$. We say that the set $\{y_k\}$ is equioscillatory if for each $s \in S$ there exists a nested neighborhood basis of 0 of radii $\epsilon(n, s)$ and for each positive integer n there exist positive integers $H(n)$ and $N(n)$ such that if $k, j \geq H(n)$ and $(y_k - y_j, P)$ is a proper pair then P contains no more than $N(n)$ points x for which $|y_k(x) - y_j(x)| > \epsilon(n, x)$.

Theorem 5.4. *Let S be a nonempty subset of R^l and $\{y_k\}$ be a sequence of functions, $y_k: S \to R^q$. The sequence $\{y_k\}$ has a subsequence which is pointwise convergent if and only if it has a subsequence $\{h_k\}$ which is pointwise bounded and $\{h_k^{(i)}\}$ is equioscillatory for $i = 1, 2, \ldots, q$.*

Proof. Apply Theorem 5.1 to $\{y_k\}$ coordinatewise to get a convergent subsequence. On the other hand if $\{y_k\}$ has a subsequence $\{h_k\}$ which converges pointwise to y, then it must be pointwise bounded. By letting $\epsilon(n, t) = \sup_{k,j\geq n}\{\|h_k(t) - h_j(t)\|\}$, $N(n) = 0$ and $H(N) = n$, then for $k, j \geq H(n)$ we have $\|h_k(t) - h_j(t)\| < \epsilon(n, t)$ for all t so that $\{h_k^{(i)}\}$ is an equioscillatory sequence for $i = 1, 2, \ldots, q$.

The following two notes relate sequences equioscillatory and sequences being Cauchy.

Note 1. The sequence $\{y_k\}$, $y_k: S \to R$, $S \subset R^l$, is equioscillatory with $N(n) = 0$ if and only if $\{y_k\}$ is pointwise Cauchy.

Proof. If $\{y_k\}$ is equioscillatory with $N(n) = 0$ then for $\delta(x) > 0$ there exists a positive integer n such that $|y_k(x) - y_j(x)| < \epsilon(n, x) < \delta(x)$ for all $k, j \geq H(n)$. If $\{y_k\}$ is pointwise Cauchy choose $\epsilon(n, x) = \sup_{k,j\geq n}\{|y_k(x) - y_j(x)|\}$, $N(n) = 0$ and $H(n) = n$ so that $\{y_k\}$ is equioscillatory.

Note 2. The sequence $\{y_k\}$, $y_k: S \to R$, $S \subset R$, is equioscillatory with $N(n) = 0$ and $\epsilon(n, x) = \epsilon_n$ a nested neighborhood basis of zero if and only if $\{y_k\}$ is uniformly Cauchy.

Proof. If $\{y_k\}$ is equioscillatory with $N(n) = 0$ and the nested neighborhood basis of 0 is $\epsilon(n, x) = \epsilon_n$ then for $\delta > 0$ there exists a positive integer n such that $\epsilon_n < \delta$ and $|y_k(x) - y_j(x)| < \epsilon_n = \epsilon(n, x)$ for all $k, j \geq H(n)$. Conversely if $\{y_k\}$ is uniformly Cauchy then for the nested neighborhood basis of radii $\epsilon(n, x) = 1/n$ there exists an $H(n)$ such that for $k, j \geq H(n)$, $|y_k(x) - y_j(x)| < 1/n = \epsilon(n, x)$ so $\{y_k\}$ is equioscillatory with $N(n) = 0$.

REFERENCES

1. P. Browne, *A singular multi-parameter eigenvalue problem in second order ordinary differential equations*, J. Differential Equations 12 (1972), 81–94.

2. P. Hartman, *Ordinary differential equations*, Wiley, New York, 1964. MR 30 #1270.

3. L. H. Hildebrandt, *Introduction to the theory of integration*, Pure and Appl. Math., vol. 13, Academic Press, New York, 1963. MR 27 #4900.

4. E. W. Hobson, *The theory of functions of a real variable and the theory of Fourier's series*, Dover, New York, 1927.

5. L. K. Jackson and K. Schrader, *Existence and uniqueness of solutions of boundary value problems for third order differential equations*, J. Differential Equations 9 (1971), 46–54. MR 42 #4813.

6. W. A. J. Luxemburg and A. C. Zaanen, *Riesz spaces*. Vol. 1, North-Holland, Amsterdam, 1971.

7. C. W. McArthur, *Convergence of monotone nets in ordered topological vector spaces*, Studia Math. 34 (1970), 1–16. MR 41 #4197.

8. E. J. McShane, *Integration*, Princeton Univ. Press, Princeton, N. J., 1944. MR 6, 43.

9. S. E. Newman, *Measure algebras and functions of bounded variation on idempotent semigroups*, Bull. Amer. Math. Soc. 75 (1969), 1396–1400. MR 40 #4778.

10. A. L. Peressini, *Ordered topological vector spaces*, Harper & Row, New York, 1967. MR 37 #3315.

11. F. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. (2) 30 (1930), 264–286.

12. ———, *The foundations of mathematics*, Harcourt-Brace, New York, 1931.

13. K. W. Schrader, *A pointwise convergence theorem for sequences of continuous functions*, Trans. Amer. Math. Soc. 159 (1971), 155–163. MR 43 #6621.

14. ———, *A generalization of the Helly selection theorem*, Bull. Amer. Math. Soc. 78 (1972), 415–419. MR 45 #8788.

15. K. Schrader and J. Thornburg, *Sufficient conditions for the existence of convergent subsequences*, Pacific J. Math. (to appear).

16. H. H. Schaefer, *Topological vector spaces*, Macmillan, New York, 1966. MR 33 #1689.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MISSOURI, COLUMBIA, MISSOURI 65201

Current address: Burroughs Corporation, 1113 North Broadway, Lexington, Kentucky 40505

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

ON THE HARISH-CHANDRA HOMOMORPHISM

BY

J. LEPOWSKY[1]

ABSTRACT. Using the Iwasawa decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}$ of a real semisimple Lie algebra $\mathfrak{g}$, Harish-Chandra has defined a now-classical homomorphism from the centralizer of $\mathfrak{k}$ in the universal enveloping algebra of $\mathfrak{g}$ into the enveloping algebra $\mathfrak{A}$ of $\mathfrak{a}$. He proved, using analysis, that its image is the space of Weyl group invariants in $\mathfrak{A}$. Here the weaker fact that the image is contained in this space of invariants is proved "purely algebraically". In fact, this proof is carried out in the general setting of semisimple symmetric Lie algebras over arbitrary fields of characteristic zero, so that Harish-Chandra's result is generalized. Related results are also obtained.

1. Introduction. Several years ago, Harish-Chandra introduced a certain mapping which now lies at the foundation of many extensive approaches to the representation theory of semisimple Lie groups. Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}$ be an Iwasawa decomposition of a real semisimple Lie algebra, $\mathfrak{G}$ the universal enveloping algebra of $\mathfrak{g}$, $\mathfrak{G}^{\mathfrak{k}}$ the centralizer of $\mathfrak{k}$ in $\mathfrak{G}$, and $\mathfrak{A}$ the universal enveloping algebra of $\mathfrak{a}$. The Harish-Chandra mapping to which we refer is the homomorphism $p: \mathfrak{G}^{\mathfrak{k}} \to \mathfrak{A}$ defined by the projection to $\mathfrak{A}$ with respect to the decomposition $\mathfrak{G} = \mathfrak{A} \oplus (\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})$ (see [2(b), §4]). Probably the most significant single property of p is that its image is contained in the algebra $\mathfrak{A}_W$ of suitably translated (by half the sum of the positive restricted roots) Weyl group invariants in $\mathfrak{A}$. (Its image actually equals $\mathfrak{A}_W$.) See [2(b), §4] for the original proof. Although this property is "purely algebraic," we know of no existing proofs which do not rely on analysis on the corresponding real semisimple Lie group. The main purpose of this paper is to present a "purely algebraic" proof of the fact that $p(\mathfrak{G}^{\mathfrak{k}}) \subset \mathfrak{A}_W$. The problem of finding such a proof was posed by B. Kostant and also by J. Dixmier.

Received by the editors January 9, 1974.

AMS (MOS) subject classifications (1970). Primary 17B20, 17B35; Secondary 17B05, 17B10, 22E45.

Key words and phrases. Harish-Chandra homomorphism, real semisimple Lie algebra, Iwasawa decomposition, universal enveloping algebra, restricted Weyl group, semisimple symmetric Lie algebra, symmetric decomposition, splitting Cartan subspace, restricted root system.

(1) Partially supported by NSF GP 33893.

In his recent book [1], Dixmier sets up an algebraic formalism which recovers the "algebraic" properties of real semisimple Lie algebras. Beginning with a "semisimple symmetric Lie algebra"—a pair $(\mathfrak{g}, \theta)$ where $\mathfrak{g}$ is a semisimple Lie algebra over an arbitrary field of characteristic zero and θ is an arbitrary automorphism of $\mathfrak{g}$ such that $\theta^2 = 1$—he obtains Cartan subspaces and Iwasawa decompositions [1, §1.13]. He then shows that certain theorems on representations of real semisimple Lie algebras, including some results of [2(a)], [5] and [6], carry over to the general context [1, Chapter 9]. Our proof of the theorem $p(\mathcal{G}^{\mathfrak{k}}) \subset \mathcal{A}_W$ holds in this setting, and therefore generalizes Harish-Chandra's original result. Our argument, which is very different from the existing analytic proofs, is not long; much of this paper consists of material on semisimple symmetric Lie algebras which is well known in the familiar special case of real semisimple Lie algebras.

The contents of this paper are as follows: In §2, we give an exposition of the relevant properties of semisimple symmetric Lie algebras, including a discussion of the restricted root system and the restricted Weyl group.

The subject of §3 is the restriction homomorphism from the algebra of $\mathfrak{k}$-invariant polynomial functions on $\mathfrak{p}$ into the algebra of polynomial functions on $\mathfrak{a}$, where $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ is the "symmetric decomposition" (i.e., eigenspace decomposition) of $\mathfrak{g}$ corresponding to θ, and $\mathfrak{a}$ is a "splitting" Cartan subspace of $\mathfrak{p}$. We show that this map injects into the algebra of Weyl group invariant polynomial functions on $\mathfrak{a}$. We do not know how to prove algebraically that it maps onto these invariants, except when $\dim \mathfrak{a} = 1$. This would be an algebraic generalization of Chevalley's polynomial restriction theorem, and could be used to prove that $p(\mathcal{G}^{\mathfrak{k}})$ is all of $\mathcal{A}_W$. Our proof of the fact that the restriction homomorphism maps into the Weyl group invariant polynomials reduces the problem to the three-dimensional simple case and solves it there. The injectivity follows from [1, Proposition 1.13.13], but our proof avoids the use of algebraic groups. We include an alternate proof, due to G. McCollum, of the key lemma for the injectivity.

The main theorem is stated and proved in §4. We include mention of the kernel of p, which is already known in the general setting (see [5, Remark 4.6], and the presentation in [1, Proposition 9.2.15]), although because of our injectivity result in §3 we can again avoid using Lie or algebraic groups. The proof that $p(\mathcal{G}^{\mathfrak{k}}) \subset \mathcal{A}_W$ is done in two stages: First we prove it when $\dim \mathfrak{a} = 1$ (and in this case we also prove that $p(\mathcal{G}^{\mathfrak{k}}) = \mathcal{A}_W$), using §3. Then we reduce the general case to this case by examining suitable semisimple subalgebras of $\mathfrak{g}$ associated with the simple restricted roots. The intermediate result, Theorem 4.17, is also interesting.

Finally, in the Appendix, we give a vector-valued generalization of the injectivity result of §3. This result also generalizes an argument in [5, proof of Lemma 4.1] (see also [6] and [1, Lemma 9.2.7, part (b) of the proof]) which depends on Lie or algebraic groups. Our original proof was simplified by G. McCollum.

We would like to thank McCollum for many valuable conversations and J. Dixmier for generously giving us access to the manuscript of his book. Certain of our methods were inspired by arguments found in [4] and [7, P. Cartier's Exposé no. 18].

Notations. The dual of a vector space V is denoted V^*. $\mathbf{Z}_+$, $\mathbf{Q}$ and $\mathbf{R}$ denote respectively the set of nonnegative integers and the fields of rational and real numbers. The restriction of a function f to a subset X of its domain is written $f|X$.

2. Preliminaries on semisimple symmetric Lie algebras. Fix a field k of characteristic zero. Let $(\mathfrak{g}, \theta)$ be a semisimple symmetric Lie algebra over k, in the sense of [1, §1.13]. That is, $\mathfrak{g}$ is a semisimple Lie algebra over k and θ is an automorphism of $\mathfrak{g}$ such that $\theta^2 = 1$. Let $\mathfrak{k}$ be the subalgebra of fixed points for θ, and let $\mathfrak{p} = \{x \in \mathfrak{g} | \theta x = -x\}$, so that $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ is a direct sum decomposition, orthogonal with respect to the Killing form of $\mathfrak{g}$. This is called the *symmetric decomposition* of $\mathfrak{g}$ defined by θ. We have $[\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p}$ and $[\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k}$.

Let $\mathfrak{a}$ be a Cartan subspace of $\mathfrak{p}$, that is, a maximal abelian subspace of $\mathfrak{p}$ which is reductive in $\mathfrak{g}$; Cartan subspaces exist by [1, Théorème 1.13.6]. Let $\mathfrak{m}$ be the centralizer of $\mathfrak{a}$ in $\mathfrak{k}$, $\mathfrak{l}$ an arbitrary Cartan subalgebra of $\mathfrak{m}$ and $\mathfrak{h} = \mathfrak{l} \oplus \mathfrak{a}$. Then $\mathfrak{h}$ is a Cartan subalgebra of $\mathfrak{g}$ [1, Proposition 1.13.7].

Let $\bar{k}$ be a field extension of k, $\bar{\mathfrak{g}} = \mathfrak{g} \otimes \bar{k}$, $\bar{\mathfrak{k}} = \mathfrak{k} \otimes \bar{k}$, etc., and let $\bar{\theta}$ be the $\bar{k}$-linear extension of θ to $\bar{\mathfrak{g}}$. Then $(\bar{\mathfrak{g}}, \bar{\theta})$ is a semisimple symmetric Lie algebra over $\bar{k}$ with symmetric decomposition $\bar{\mathfrak{g}} = \bar{\mathfrak{k}} \oplus \bar{\mathfrak{p}}$, $\bar{\mathfrak{a}}$ is a Cartan subspace of $\bar{\mathfrak{p}}$, $\bar{\mathfrak{m}}$ is the centralizer of $\bar{\mathfrak{a}}$ in $\bar{\mathfrak{k}}$, $\bar{\mathfrak{l}}$ (resp., $\bar{\mathfrak{h}}$) is a Cartan subalgebra of $\bar{\mathfrak{m}}$ (resp., $\bar{\mathfrak{g}}$), and $\bar{\mathfrak{h}} = \bar{\mathfrak{l}} \oplus \bar{\mathfrak{a}}$.

Suppose that $\bar{\mathfrak{h}}$ is a splitting Cartan subalgebra of $\bar{\mathfrak{g}}$. (This can be insured by choosing $\bar{k}$ to be algebraically closed.) Denote by $R \subset \bar{\mathfrak{h}}^*$ the set of roots of $\bar{\mathfrak{g}}$ with respect to $\bar{\mathfrak{h}}$.

Assume now that $\mathfrak{a}$ is a splitting Cartan subspace of $\mathfrak{p}$ in the sense of [1, §1.13], i.e., for all $a \in \mathfrak{a}$, the operator ad a on $\mathfrak{g}$ can be upper triangularized and hence diagonalized. Consider the root space decomposition of $\bar{\mathfrak{g}}$ with respect to $\bar{\mathfrak{h}}$:

$$\bar{\mathfrak{g}} = \bar{\mathfrak{h}} \oplus \coprod_{\lambda \in R} \bar{\mathfrak{g}}^{\lambda},$$

where $\bar{\mathfrak{g}}^\lambda$ denotes the root space for λ. Let $P: \bar{\mathfrak{h}}^* \to \bar{\mathfrak{a}}^*$ denote the restriction map, and let Σ denote the set of nonzero members of $P(R)$. For all $\bar{k}$-linear functionals $\phi: \bar{\mathfrak{a}} \to \bar{k}$, let

$$\bar{\mathfrak{g}}^\phi = \{x \in \bar{\mathfrak{g}} | [a, x] = \phi(a)x \text{ for all } a \in \bar{\mathfrak{a}}\}.$$

Then clearly

$$\bar{\mathfrak{g}}^0 = \bar{\mathfrak{m}} \oplus \bar{\mathfrak{a}}, \quad \text{and} \quad \bar{\mathfrak{g}} = \bar{\mathfrak{g}}^0 \oplus \coprod_{\phi \in \Sigma} \bar{\mathfrak{g}}^\phi.$$

Now for all k-linear functionals $\phi: \mathfrak{a} \to k$, define

$$\mathfrak{g}^\phi = \{x \in \mathfrak{g} | [a, x] = \phi(a)x \text{ for all } a \in \mathfrak{a}\}.$$

Then

$$\mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a},$$

and since $\mathfrak{a}$ is a splitting Cartan subspace, Σ is identified with (i.e., is the set of $\bar{k}$-linear extensions of the members of) $\{\phi \in \mathfrak{a}^* | \phi \neq 0 \text{ and } \mathfrak{g}^\phi \neq 0\}$, and

$$\mathfrak{g} = \mathfrak{g}^0 \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^\phi = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^\phi.$$

The members of Σ, regarded as elements of either $\mathfrak{a}^*$ or $\bar{\mathfrak{a}}^*$, are called the *restricted roots* of $\mathfrak{g}$ with respect to $\mathfrak{a}$. Σ spans $\mathfrak{a}^*$ over k and $\bar{\mathfrak{a}}^*$ over $\bar{k}$. Note that $[\mathfrak{g}^\phi, \mathfrak{g}^\psi] \subset \mathfrak{g}^{\phi+\psi}$ and that $\theta\mathfrak{g}^\phi = \mathfrak{g}^{-\phi}$ for all $\phi, \psi \in \mathfrak{a}^*$.

For all $\phi \in \Sigma$, $\bar{\mathfrak{g}}^\phi = \coprod \bar{\mathfrak{g}}^\lambda$, where λ ranges through $\{\lambda \in R | P(\lambda) = \phi\}$, and setting $R' = \{\lambda \in R | P(\lambda) = 0\}$, we have

$$\bar{\mathfrak{g}}^0 = \bar{\mathfrak{h}} \oplus \coprod_{\lambda \in R'} \bar{\mathfrak{g}}^\lambda.$$

Moreover, $R'|\bar{\mathfrak{l}}$ is the set of roots of the reductive Lie algebra $\bar{\mathfrak{m}}$ with respect to its splitting Cartan subalgebra $\bar{\mathfrak{l}}$; for all $\lambda \in R'$, the root space $\bar{\mathfrak{m}}^{\lambda|\bar{\mathfrak{l}}} = \bar{\mathfrak{g}}^\lambda$; and

$$\bar{\mathfrak{m}} = \bar{\mathfrak{l}} \oplus \coprod_{\lambda \in R'} \bar{\mathfrak{g}}^\lambda$$

(see [1, Propositions 1.13.7(iii) and 1.13.9]). Also, setting $R'' = \{\lambda \in R | P(\lambda) \neq 0\}$, we have $P(R'') = \Sigma$ by definition.

Let B be the Killing form of $\mathfrak{g}$ and $\bar{B}$ its $\bar{k}$-bilinear extension to $\bar{\mathfrak{g}}$, so that $\bar{B}$ is the Killing form of $\bar{\mathfrak{g}}$. Then $\bar{B}$ is nonsingular on $\bar{\mathfrak{h}}$, and so defines naturally a nonsingular symmetric bilinear form, denoted $\bar{B}^*$, on $\bar{\mathfrak{h}}^*$. Moreover, since $\bar{B}(\bar{\mathfrak{l}}, \bar{\mathfrak{a}}) = 0$, $\bar{B}$ is nonsingular on $\bar{\mathfrak{a}}$, thus defining a nonsingular symmetric bilinear form on $\bar{\mathfrak{a}}^*$. Let us identify $\bar{\mathfrak{a}}^*$ with the subspace $\{\lambda \in \bar{\mathfrak{h}}^* | \lambda | \bar{\mathfrak{l}} = 0\}$ of $\bar{\mathfrak{h}}^*$ by extending the definition of each element of $\bar{\mathfrak{a}}^*$ by requiring it to be zero on $\bar{\mathfrak{l}}$. Then the natural bilinear form on $\bar{\mathfrak{a}}^*$ is exactly the restriction of $\bar{B}^*$ to $\bar{\mathfrak{a}}^*$. Moreover, if we also identify $\bar{\mathfrak{l}}^*$ with the sub-

space of elements of $\overline{\mathfrak{h}}^*$ vanishing on $\overline{\mathfrak{a}}$, then $\overline{\mathfrak{h}}^* = \overline{\mathfrak{l}}^* \oplus \overline{\mathfrak{a}}^*$, and $\overline{B}^*(\overline{\mathfrak{l}}^*, \overline{\mathfrak{a}}^*) = 0$. In particular, the restriction map $P : \overline{\mathfrak{h}}^* \to \overline{\mathfrak{a}}^*$ coincides both with the projection to $\overline{\mathfrak{a}}^*$ with respect to the above decomposition and with the orthogonal projection to $\overline{\mathfrak{a}}^*$ with respect to $\overline{B}^*$.

The automorphism $\overline{\theta}$ of $\overline{\mathfrak{g}}$ is 1 on $\overline{\mathfrak{l}}$ and -1 on $\overline{\mathfrak{a}}$, and so preserves $\overline{\mathfrak{h}}$. The transpose of $\overline{\theta}|\overline{\mathfrak{h}}$, which we denote by $\overline{\theta}^*$, is the isometry of $\overline{\mathfrak{h}}^*$ which is 1 on $\overline{\mathfrak{l}}^*$ and -1 on $\overline{\mathfrak{a}}^*$. Thus $P : \overline{\mathfrak{h}}^* \to \overline{\mathfrak{a}}^*$ can be realized by the formula $\frac{1}{2}(1 - \overline{\theta}^*)$.

Now let $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$ denote the rational span of R in $\overline{\mathfrak{h}}^*$. Then $\overline{\mathfrak{h}}^* = \overline{\mathfrak{h}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \overline{k}$ in a natural way. Moreover, $\overline{B}^*$ is $\mathbf{Q}$-valued and positive definite on the rational space $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$. Now $\overline{\theta}^*$ preserves R and hence preserves $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$, and so $P = \frac{1}{2}(1 - \overline{\theta}^*)$ also preserves $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$. Thus since Σ consists of the nonzero members of $P(R)$, $\Sigma \subset \overline{\mathfrak{h}}^*_{\mathbf{Q}}$, so that for all $\phi \in \Sigma$, $\overline{B}^*(\phi, \phi)$ is a positive rational number. But B is nonsingular on $\mathfrak{a}$ because $\overline{B}$ is nonsingular on $\overline{\mathfrak{a}}$. Hence B induces naturally a nonsingular symmetric k-bilinear form $(\cdot, \cdot)$ on $\mathfrak{a}^*$. If we identify $\overline{\mathfrak{a}}^*$ with $\mathfrak{a}^* \otimes_k \overline{k}$, then $\overline{B}^*|\mathfrak{a}^*$ is just $(\cdot, \cdot)$ since $\overline{B}^*|\overline{\mathfrak{a}}^*$ is the canonical form on $\overline{\mathfrak{a}}^*$ defined by $\overline{B}|\overline{\mathfrak{a}}$. Hence $\overline{B}^*(\phi, \phi) = (\phi, \phi)$ for all $\phi \in \Sigma$, and we have the following two lemmas:

Lemma 2.1. *Let* $\overline{\mathfrak{a}}^*_{\mathbf{Q}} = \overline{\mathfrak{a}}^* \cap \overline{\mathfrak{h}}^*_{\mathbf{Q}}$ *and* $\overline{\mathfrak{l}}^*_{\mathbf{Q}} = \overline{\mathfrak{l}}^* \cap \overline{\mathfrak{h}}^*_{\mathbf{Q}}$. *Then* $\overline{\mathfrak{h}}^*_{\mathbf{Q}} = \overline{\mathfrak{a}}^*_{\mathbf{Q}} \oplus \overline{\mathfrak{l}}^*_{\mathbf{Q}}$, $\overline{\mathfrak{h}}^* = \overline{\mathfrak{h}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \overline{k}$, $\overline{\mathfrak{a}}^* = \overline{\mathfrak{a}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \overline{k}$ *and* $\overline{\mathfrak{l}}^* = \overline{\mathfrak{l}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \overline{k}$. *Moreover,* $\overline{\mathfrak{a}}^*_{\mathbf{Q}}$ *is the rational span of* Σ, $\overline{\mathfrak{a}}^*_{\mathbf{Q}} \subset \mathfrak{a}^*$ *and* $\mathfrak{a}^* = \overline{\mathfrak{a}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} k$.

Lemma 2.2. $\overline{B}^*|\overline{\mathfrak{h}}^*_{\mathbf{Q}}$ *is rational-valued and positive definite.* B *is nonsingular on* $\mathfrak{a}$. *Denoting by* $(\cdot, \cdot)$ *the corresponding canonical nonsingular symmetric k-bilinear form on* $\mathfrak{a}^*$, *we have that* $\overline{B}^*|\mathfrak{a}^* = (\cdot, \cdot)$. *In particular,* $(\cdot, \cdot)$ *is rational-valued and positive definite on* $\overline{\mathfrak{a}}^*_{\mathbf{Q}}$, *and for all* $\phi \in \Sigma, (\phi, \phi)$ *is a positive rational number.*

For each $\lambda \in R$, let $w_\lambda : \overline{\mathfrak{h}}^* \to \overline{\mathfrak{h}}^*$ denote the orthogonal reflection through the hyperplane of $\overline{\mathfrak{h}}^*$ orthogonal to λ (with respect to $\overline{B}^*$). Then w_λ is an isometry of $\overline{\mathfrak{h}}^*$ which preserves $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$ and R. The *Weyl group* W_R of $\overline{\mathfrak{g}}$ with respect to $\overline{\mathfrak{h}}$ is defined to be the group of isometries of $\overline{\mathfrak{h}}^*$ or of $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$ generated by the w_λ ($\lambda \in R$).

Now let E be the real vector space $\overline{\mathfrak{h}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \mathbf{R}$. Then the restriction of $\overline{B}^*$ to $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$ extends naturally to an $\mathbf{R}$-bilinear form B_E on E. Since $\overline{B}^*$ is positive definite on $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$, B_E is positive definite on E and hence is a Euclidean scalar product on E. W_R extends naturally to a group of isometries, also denoted W_R, of E, and R becomes a reduced system of roots in E with Weyl group W_R, in the sense of [1, Appendice] and [8, §1.1.2].

Recall that the isometry $\overline{\theta}^*$ of $\overline{\mathfrak{h}}^*$ preserves R and $\overline{\mathfrak{h}}^*_{\mathbf{Q}}$. The $\mathbf{R}$-linear extension of $\overline{\theta}^*|\overline{\mathfrak{h}}^*_{\mathbf{Q}}$ to E is an isometry of E with square 1 which we call

θ_E. Let $\sigma = -\theta_E$, so that σ is an isometry of E which preserves R and has square 1. Then (R, σ) is a σ-system of roots in E, in the sense of [8, §1.1.3], except that we allow the cases $\sigma = \pm 1$.

Let E_+ (resp., E_-) denote the $+1$ (resp., -1) eigenspace of σ in E, so that $E = E_+ \oplus E_-$, and this is an orthogonal deceomposition. Then

$$\overline{\mathfrak{a}}^*_{\mathbf{Q}} = E_+ \cap \overline{\mathfrak{h}}^*_{\mathbf{Q}}, \quad \overline{\mathfrak{t}}^*_{\mathbf{Q}} = E_- \cap \overline{\mathfrak{h}}^*_{\mathbf{Q}}, \quad E_+ = \overline{\mathfrak{a}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \mathbf{R}, \quad E_- = \overline{\mathfrak{t}}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} \mathbf{R},$$

and E_+ is the real span of Σ. Recall that we have defined $R' = \{\lambda \in R | P(R) = 0\}$ and $R'' = \{\lambda \in R | P(R) \neq 0\}$. Then clearly $R' = R \cap E_-$,

$$R' = \{\lambda \in R | \sigma\lambda = -\lambda\} = \{\lambda \in R | \lambda|\overline{\mathfrak{a}} = 0\}$$

and

$$R'' = \{\lambda \in R | \sigma\lambda \neq -\lambda\} = \{\lambda \in R | \lambda|\overline{\mathfrak{a}} \neq 0\}.$$

Recall that if $\lambda \in R'$, then the root space $\overline{\mathfrak{g}}^{\lambda} \subset \overline{\mathfrak{m}}$.

Lemma 2.3. *For all* $\lambda \in R$, $\sigma\lambda - \lambda \notin R$.

Proof. Assume that $\sigma\lambda - \lambda \in R$. Then $\sigma\lambda - \lambda \in R'$, and so $\lambda - \sigma\lambda \in R'$. Thus $\overline{\mathfrak{g}}^{\lambda-\sigma\lambda} \subset \overline{\mathfrak{m}}$. Let e_λ be a nonzero vector in $\overline{\mathfrak{g}}^{\lambda}$. Then $\overline{\theta} e_\lambda \in \overline{\mathfrak{g}}^{\overline{\theta}^*\lambda} = \overline{\mathfrak{g}}^{-\sigma\lambda}$, and so $[e_\lambda, \overline{\theta} e_\lambda]$ is a nonzero element of $\overline{\mathfrak{g}}^{\lambda-\sigma\lambda} \subset \overline{\mathfrak{m}}$. But $\overline{\theta}[e_\lambda, \overline{\theta} e_\lambda] = [\overline{\theta} e_\lambda, e_\lambda] = -[e_\lambda, \overline{\theta} e_\lambda]$, so that $[e_\lambda, \overline{\theta} e_\lambda] \in \overline{\mathfrak{p}}$. This contradiction proves the lemma. Q.E.D.

This result asserts that (R, σ) is a normal σ-system, in the sense of [8, §1.1.3]. (The proof here, which also appears in [3, p. 76, proof of Lemma 3.6], is more direct than the proof given in [8, Lemma 1.1.3.6] for the special case of real semisimple Lie algebras.) The results of [8, §1.1.3] on normal σ-systems are now applicable in our context. (The cases $\sigma = \pm 1$ do not cause any difficulty.) In particular, since Σ is the set of orthogonal projections to E_+ of the members of R'', we have by [8, Proposition 1.1.3.1]:

Lemma 2.4 (S. Araki). *Σ is a (not necessarily reduced) system of roots in E_+, in the sense of* [8, §1.1.2]. *The Weyl group W of Σ is the group of isometries of E_+ (with respect to B_E) generated by $\{s_\phi | \phi \in \Sigma\}$ where s_ϕ is the reflection through the hyperplane of E_+ orthogonal to ϕ.*

W preserves $E_+ \cap \overline{\mathfrak{h}}^*_{\mathbf{Q}} = \overline{\mathfrak{a}}^*_{\mathbf{Q}}$, and its restriction to this space extends naturally to a group of isometries, still denoted W, of $\mathfrak{a}^*$, with respect to the bilinear form $(\cdot,\cdot)$ in Lemma 2.2. In this context, W is called the *restricted Weyl group* of $\mathfrak{g}$ with respect to $\mathfrak{a}$. It is the group of isometries of $\mathfrak{a}^*$ (with respect to $(\cdot,\cdot)$) generated by $\{s_\phi | \phi \in \Sigma\}$, where in this case s_ϕ is identified with the orthogonal reflection through the hyperplane of $\mathfrak{a}^*$ orthogonal to

ϕ. Similarly, W extends to a group of isometries of $\overline{\mathfrak{a}}^*$.

We shall need the following result:

Lemma 2.5. *Let* $\phi \in \Sigma$ *and* $s \in W$. *Then* $\dim \mathfrak{g}^{\phi} = \dim \mathfrak{g}^{s\phi}$.

Proof. It is sufficient to show that the $\overline{k}$-dimensions of $\overline{\mathfrak{g}}^{\phi}$ and $\overline{\mathfrak{g}}^{s\phi}$ are equal. By [8, Lemma 1.1.3.5 or Proposition 1.1.3.3], there exists $w \in W_R$ (the Weyl group for $\overline{\mathfrak{g}}$ with respect to $\overline{\mathfrak{h}}$) such that w, regarded as an isometry of $\overline{\mathfrak{h}}^*$, preserves $\overline{\mathfrak{a}}^*$ and $\overline{\mathfrak{l}}^*$, and restricts to s on $\overline{\mathfrak{a}}^*$. Hence $P \circ w = w \circ s : \overline{\mathfrak{h}}^* \to \overline{\mathfrak{a}}^*$. But w preserves R, and hence takes $R_{\phi} = \{\lambda \in R \mid P(\lambda) = \phi\}$ onto $R_{s\phi} = \{\lambda \in R \mid P(\lambda) = s\phi\}$. Since

$$\overline{\mathfrak{g}}^{\phi} = \coprod_{\lambda \in R_{\phi}} \overline{\mathfrak{g}}^{\lambda}, \quad \overline{\mathfrak{g}}^{s\phi} = \coprod_{\lambda \in R_{s\phi}} \overline{\mathfrak{g}}^{\lambda},$$

and each $\overline{\mathfrak{g}}^{\lambda}$ is one-dimensional, the lemma is proved. Q.E.D.

Let Σ_+ be an arbitrary positive system in the root system Σ. Then we have the decomposition

$$(*) \qquad \mathfrak{g} = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma_+} \mathfrak{g}^{\phi} \oplus \coprod_{\phi \in \Sigma_+} \mathfrak{g}^{-\phi}.$$

Let $\mathfrak{n}$ be the subalgebra $\coprod \mathfrak{g}^{\phi}$ $(\phi \in \Sigma_+)$ of $\mathfrak{g}$.

Lemma 2.6. *We have* $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}$.

Proof. To show that $\mathfrak{g} = \mathfrak{k} + \mathfrak{a} + \mathfrak{n}$, it is sufficient, in view of $(*)$, to show that $\mathfrak{g}^{-\phi} \subset \mathfrak{k} + \mathfrak{n}$ for all $\phi \in \Sigma_+$. But if $x \in \mathfrak{g}^{-\phi}$, $\theta x \in \theta \mathfrak{g}^{-\phi} = \mathfrak{g}^{\phi}$, so that $x = (x + \theta x) - \theta x \in \mathfrak{k} + \mathfrak{n}$. Now suppose $x \in \mathfrak{k}$, $y \in \mathfrak{a}$ and $z \in \mathfrak{n}$, and suppose $x + y + z = 0$. Then $0 = \theta(x + y + z) = x - y + \theta z$, so that $2y + z - \theta z = 0$. But $\theta z \in \coprod_{\phi \in \Sigma_+} \mathfrak{g}^{-\phi}$, and so by the directness of the sum $(*)$, $y = z = 0$. Hence $x = 0$ also. Q.E.D.

This decomposition is called the *Iwasawa decomposition* of $\mathfrak{g}$ associated with θ, $\mathfrak{a}$ and Σ_+. We may choose a positive system R_+ for R such that $\{\lambda \in R \mid P(\lambda) \in \Sigma_+\} = R''_+$, where $R''_+ = R'' \cap R_+$, and the Iwasawa decomposition implies that $\overline{\mathfrak{g}} = \overline{\mathfrak{k}} \oplus \overline{\mathfrak{a}} \oplus \coprod_{\phi \in R''_+} \overline{\mathfrak{g}}^{\phi}$. We shall not need this last fact.

3. The polynomial restriction map F_*. Let k be a field of characteristic zero. For every finite-dimensional vector space V over k, let $S(V)$ denote the symmetric algebra over V. Since k is infinite, $S(V^*)$ is naturally isomorphic to the algebra of polynomial functions on V, i.e., the algebra of sums of products of linear functionals on V. (The isomorphism takes the symmetric algebra product $f_1 f_2 \cdots f_r$ of linear functionals f_i to the corresponding product $f_1 f_2 \cdots f_r$ of functions on V.) Hence we may, and often shall, identify

$S(V^*)$ with the algebra of polynomial functions on V.

Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the symmetric decomposition of a semisimple symmetric Lie algebra $(\mathfrak{g}, \theta)$ over k, $\mathfrak{a} \subset \mathfrak{p}$ a splitting Cartan subspace, and W the corresponding restricted Weyl group. Then $\mathfrak{k}$ acts on $\mathfrak{p}$, hence on $\mathfrak{p}^*$ by contragredience, and thus on $S(\mathfrak{p}^*)$ by unique extension by derivations. Denote by $S(\mathfrak{p}^*)^{\mathfrak{k}}$ the corresponding algebra of $\mathfrak{k}$-annihilated vectors. Also, W acts on $\mathfrak{a}^*$, and hence acts on $S(\mathfrak{a}^*)$ by automorphisms. Let $S(\mathfrak{a}^*)^W$ be the algebra of W-invariants.

Denote by $F : S(\mathfrak{p}^*) \to S(\mathfrak{a}^*)$ the restriction homomorphism, and let $F_* = F|S(\mathfrak{p}^*)^{\mathfrak{k}}$, so that $F_* : S(\mathfrak{p}^*)^{\mathfrak{k}} \to S(\mathfrak{a}^*)$.

Theorem 3.1. *We have $F_*(S(\mathfrak{p}^*)^{\mathfrak{k}}) \subset S(\mathfrak{a}^*)^W$, and $F_* : S(\mathfrak{p}^*)^{\mathfrak{k}} \to S(\mathfrak{a}^*)^W$ is an algebra injection.*

Proof. By passing to an algebraic closure of k if necessary, we observe that to prove the theorem it is sufficient to prove it in case k is algebraically closed. We now make this assumption. However, we shall need it only in the proof that $F_*(S(\mathfrak{p}^*)^{\mathfrak{k}}) \subset S(\mathfrak{a}^*)^W$, and only after Lemma 3.3.

First we shall show that $F_*(S(\mathfrak{p}^*)^{\mathfrak{k}}) \subset S(\mathfrak{a}^*)^W$ and then that F_* is injective. The first assertion will be proved essentially by reducing to the case in which $\mathfrak{g}$ is three-dimensional simple.

Let B be the Killing form of $\mathfrak{g}$. Define a bilinear form B_θ on $\mathfrak{g}$ by

$$B_\theta(x, y) = -B(x, \theta y)$$

for all $x, y \in \mathfrak{g}$, so that B_θ is a nonsingular symmetric form.

Let $\Sigma \subset \mathfrak{a}^*$ be the set of restricted roots of $\mathfrak{g}$ with respect to $\mathfrak{a}$, and let $\mathfrak{m}$ be the centralizer of $\mathfrak{a}$ in $\mathfrak{k}$.

Lemma 3.2. *The decomposition $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^\phi$ is a B_θ-orthogonal decomposition. In particular, B_θ is nonsingular on each $\mathfrak{g}^\phi$ $(\phi \in \Sigma)$, on $\mathfrak{m}$ and on $\mathfrak{a}$. Also, B is nonsingular on $\mathfrak{m}$ and on $\mathfrak{a}$.*

Proof. It is easy to see that for all $\phi, \psi \in \Sigma$, $B(\mathfrak{g}^\phi, \mathfrak{g}^\psi) = 0$ unless $\phi + \psi = 0$, and that $B(\mathfrak{g}^\phi, \mathfrak{m} + \mathfrak{a}) = 0$. Since $\theta \mathfrak{g}^\phi = \mathfrak{g}^{-\phi}$ for all $\phi \in \Sigma$ and since $B_\theta(\mathfrak{m}, \mathfrak{a}) = B(\mathfrak{m}, \mathfrak{a}) = 0$, the decomposition $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^\phi$ is B_θ-orthogonal. Since B_θ is nonsingular, the restriction of B_θ to each component in this sum is nonsingular. Q.E.D.

(The nonsingularity of B on $\mathfrak{a}$ was proved another way in §2; see Lemma 2.2.)

Since B is nonsingular on $\mathfrak{a}$, B induces a nonsingular symmetric bilinear form $(\cdot,\cdot)$ on $\mathfrak{a}^*$ and an isometry from $\mathfrak{a}^*$ onto $\mathfrak{a}$. For all $\phi \in \Sigma$, let

$x_\phi \in \mathfrak{a}$ be the image of ϕ under this isometry. Then $B(x_\phi, a) = \phi(a)$ for all $a \in \mathfrak{a}$, and $B(x_\phi, x_\psi) = \phi(x_\psi) = (\phi, \psi)$ for all $\phi, \psi \in \Sigma$.

Lemma 3.3. *Let* $\phi \in \Sigma$. *For all* $e \in \mathfrak{g}^\phi$,

$$[e, \theta e] = B(e, \theta e)x_\phi = -B_\theta(e, e)x_\phi.$$

Proof. Since $\theta e \in \theta \mathfrak{g}^\phi = \mathfrak{g}^{-\phi}$, $[e, \theta e] \in \mathfrak{m} + \mathfrak{a}$. But $\theta[e, \theta e] = -[e, \theta e]$, so that $[e, \theta e] \in \mathfrak{p}$. Hence $[e, \theta e] \in \mathfrak{a}$. Since B is nonsingular on $\mathfrak{a}$, it is sufficient to show that $B(a, [e, \theta e]) = B(a, B(e, \theta e)x_\phi)$ for all $a \in \mathfrak{a}$. But

$$B(a, [e, \theta e]) = B([a, e], \theta e) = B(\phi(a)e, \theta e) = \phi(a)B(e, \theta e) = B(a, x_\phi)B(e, \theta e)$$

$$= B(a, B(e, \theta e)x_\phi),$$

proving the lemma. Q.E.D.

Let $\phi \in \Sigma$. Since $(\phi, \phi) \neq 0$ (see Lemma 2.2), we can define

$$h_\phi = 2x_\phi/(\phi, \phi) \in \mathfrak{a},$$

and we have $\phi(h_\phi) = 2$.

Also, since B_θ is a symmetric nonsingular form on $\mathfrak{g}^\phi$ (Lemma 3.2), there exists $e \in \mathfrak{g}^\phi$ such that $B_\theta(e, e) \neq 0$. Setting

$$e_\phi = (2/(\phi, \phi)B_\theta(e, e))^{1/2}e,$$

which we may do since k is algebraically closed, we get

$$B_\theta(e_\phi, e_\phi) = 2/(\phi, \phi).$$

Thus from Lemma 3.3, we have:

Lemma 3.4. $[h_\phi, e_\phi] = 2e_\phi$, $[h_\phi, -\theta e_\phi] = 2\theta e_\phi$ *and* $[e_\phi, -\theta e_\phi] = h_\phi$. *In particular,* $\{h_\phi, e_\phi, \theta e_\phi\}$ *spans a three-dimensional simple subalgebra* $\mathfrak{g}_\phi$ *of* $\mathfrak{g}$.

It is clear that $\mathfrak{g}_\phi$ is stable under θ, so that $\mathfrak{g}_\phi = \mathfrak{k}_\phi \oplus \mathfrak{p}_\phi$, where $\mathfrak{k}_\phi = \mathfrak{g}_\phi \cap \mathfrak{k}$ and $\mathfrak{p}_\phi = \mathfrak{g}_\phi \cap \mathfrak{p}$. Moreover, $(\mathfrak{g}_\phi, \theta|\mathfrak{g}_\phi)$ is a semisimple symmetric Lie algebra. We have

$$\mathfrak{k}_\phi = k(e_\phi + \theta e_\phi) \quad \text{and} \quad \mathfrak{p}_\phi = kh_\phi \oplus k(e_\phi - \theta e_\phi).$$

We shall use $\mathfrak{g}_\phi$ to show that the restriction to $\mathfrak{a}$ of every element of $S(\mathfrak{p}^*)^\mathfrak{k}$ is invariant under the Weyl reflection with respect to ϕ.

From Lemma 3.4, we have

$$[½(e_\phi + \theta e_\phi), h_\phi] = -(e_\phi - \theta e_\phi), \quad [½(e_\phi + \theta e_\phi), e_\phi - \theta e_\phi] = h_\phi$$

and $[\frac{1}{2}(e_\phi + \theta e_\phi), \mathfrak{t}] = 0$, where $\mathfrak{t} = \operatorname{Ker} \phi \subset \mathfrak{a}$.

For all $f \in S(\mathfrak{p}^*)$, $(\frac{1}{2}(e_\phi + \theta e_\phi) \cdot f)|(\mathfrak{p}_\phi \oplus \mathfrak{t}) = \frac{1}{2}(e_\phi + \theta e_\phi) \cdot (f|\mathfrak{p}_\phi \oplus \mathfrak{t})$. In particular, if $f \in S(\mathfrak{p}^*)^{\mathfrak{k}}$, then $f' = f|(\mathfrak{p}_\phi \oplus \mathfrak{t})$ is a $\mathfrak{k}_\phi$-annihilated polynomial function on $\mathfrak{p}_\phi \oplus \mathfrak{t}$, where $\mathfrak{k}_\phi = k(e_\phi \oplus \theta e_\phi)$ acts on $\mathfrak{p}_\phi \oplus \mathfrak{t}$ as indicated above. The determination of f' is a simple, standard problem in "classical invariant theory," which we proceed to solve.

Let $x = h_\phi + (-1)^{1/2}(e_\phi - \theta e_\phi)$, $y = h_\phi - (-1)^{1/2}(e_\phi - \theta e_\phi)$ and let $\{z_1, \ldots, z_n\}$ be a basis of $\mathfrak{t}$. Then the basis $x, y, z_1, \ldots, z_n$ of $\mathfrak{p}_\phi \oplus \mathfrak{t} = \mathfrak{p}_\phi + \mathfrak{a}$ diagonalizes the action of $\frac{1}{2}(-1)^{1/2}(e_\phi + \theta e_\phi)$, with eigenvalues -1, $1, 0, \ldots, 0$, respectively.

Let $X, Y, Z_1, \ldots, Z_n$ be the basis of $(\mathfrak{p}_\phi \oplus \mathfrak{t})^*$ dual to $x, y, z_1, \ldots, z_n$, so that f' is a polynomial in these variables. Moreover, $\frac{1}{2}(-1)^{1/2}(e_\phi + \theta e_\phi)$ acts on such a polynomial by the derivation law and the negative transpose of the action on $\mathfrak{p}_\phi \oplus \mathfrak{t}$. Hence $X, Y, Z_1, \ldots, Z_n$ is a basis of eigenvectors of $(\mathfrak{p}_\phi \oplus \mathfrak{t})^*$ with eigenvalues $1, -1, 0, \ldots, 0$, respectively.

Write

$$f' = \sum_{j,k \in \mathbf{Z}_+} X^j Y^k f_{jk}(Z_1, \ldots, Z_n),$$

where the f_{jk}'s are uniquely determined polynomials in the Z_i's. The invariance of f' under $\frac{1}{2}(-1)^{1/2}(e_\phi + \theta e_\phi)$ asserts that

$$\sum_{j,k \in \mathbf{Z}_+} (j-k) X^j Y^k f_{jk}(Z_1, \ldots, Z_n) = 0,$$

i.e., that $f_{jk} = 0$ for all pairs j, k such that $j \neq k$. Hence f' is $\frac{1}{2}(-1)^{1/2}(e_\phi + \theta e_\phi)$-invariant if and only if f' is of the form

$$f' = \sum_{j \in \mathbf{Z}_+} (XY)^j f_j(Z_1, \ldots, Z_n),$$

which the f_j's are polynomials in the Z_i's.

Now let $H = X + Y$ and $E = (-1)^{1/2}(X - Y)$, so that the basis $H, E, Z_1, \ldots, Z_n$ of $(\mathfrak{p}_\phi \oplus \mathfrak{t})^*$ is dual to the basis $h_\phi, e_\phi - \theta e_\phi, z_1, \ldots, z_n$ of $\mathfrak{p}_\phi \oplus \mathfrak{t}$. Then f' is of the form

$$f' = \sum_{j \in \mathbf{Z}_+} (H^2 + E^2)^j g_j(Z_1, \ldots, Z_n),$$

where the g_j's are polynomials in the Z_i's.

Since $H|\mathfrak{a}$ is a nonzero multiple of ϕ, $E|\mathfrak{a} = 0$ and each $Z_i|\mathfrak{a}$ annihilates h_ϕ, we have

$$f'|\mathfrak{a} = \sum_{j \in \mathbf{Z}_+} \phi^{2j} h_j,$$

where each h_j is a polynomial in linear functionals on $\mathfrak{a}$ which are orthogonal to ϕ with respect to $(\cdot,\cdot)$. We have shown that for all $f \in S(\mathfrak{p}^*)^{\mathfrak{k}}$, $f|\mathfrak{a} = f'|\mathfrak{a}$ is an element of $S(\mathfrak{a}^*)$ left fixed by the Weyl reflection with respect to ϕ. Since these reflections generate W as ϕ ranges through Σ, $f|\mathfrak{a} \in S(\mathfrak{a}^*)^W$. This proves the first part of the theorem.

We turn now to the injectivity of F_*. In the following proof, we need not assume that k is algebraically closed. We begin with some general comments on symmetric algebras.

Let V be a finite-dimensional vector space (over k), and for all $r \in \mathbf{Z}_+$, let $S^r(V)$ denote the rth homogeneous component of $S(V)$. There is a natural pairing $\{\cdot,\cdot\}$ between $S^r(V^*)$ and $S^r(V)$ given as follows:

$$\{f_1 \cdots f_r, v_1 \cdots v_r\} = \sum_{\sigma} \prod_{i=1}^{r} \langle f_i, v_{\sigma(i)}\rangle,$$

where $v_1, \ldots, v_r \in V$, $f_1, \ldots, f_r \in V^*$, $\langle\cdot,\cdot\rangle$ is the natural pairing between V^* and V and σ ranges through the group of permutations of $\{1, \ldots, r\}$. Then for all $v \in V$ and $f \in S^r(V^*)$, $\{f, v^r\} = r!f(v)$ (regarding f as a polynomial function on V on the right-hand side). We have two immediate consequences:

Lemma 3.5. (i) $\{\cdot,\cdot\}$ *is a nonsingular pairing.*

(ii) *Let Z be a Zariski dense subset of V (i.e., for all $f \in S(V^*)$, $f(Z) = 0$ implies $f = 0$). Then $\{z^r | z \in Z\}$ spans $S^r(V)$.*

Now $\mathfrak{k}$ acts naturally as derivations on $S(\mathfrak{p})$ and $S(\mathfrak{p}^*)$.

Lemma 3.6. *For each $r \in \mathbf{Z}_+$, the natural actions of $\mathfrak{k}$ on $S^r(\mathfrak{p}^*)$ and $S^r(\mathfrak{p})$ are contragredient under $\{\cdot,\cdot\}$.*

Proof. Let $x \in \mathfrak{k}$, $y \in \mathfrak{p}$ and $z \in \mathfrak{p}^*$. By Lemma 3.5(ii), it is sufficient to show that $\{x \cdot z^r, y^r\} = -\{z^r, x \cdot y^r\}$. But

$$\{x \cdot z^r, y^r\} = r\{(x \cdot z)z^{r-1}, y^r\} = rr!\langle x \cdot z, y\rangle\langle z, y\rangle^{r-1}$$

$$= -rr!\langle z, [x, y]\rangle\langle z, y\rangle^{r-1} = -r\{z^r, [x, y]y^{r-1}\} = -\{z^r, x \cdot y^r\},$$

and this proves the lemma. Q.E.D.

The next lemma is the crucial one. We give two proofs, the first inspired by P. Cartier's argument in [7, p. 18–20, Proposition 1], and the second due to G. McCollum.

Lemma 3.7. *Under the natural action of $\mathfrak{k}$ on $S^r(\mathfrak{p})$, $S^r(\mathfrak{a})$ generates $S^r(\mathfrak{p})$. In particular,*

$$S^r(\mathfrak{p}) = S^r(\mathfrak{a}) + \mathfrak{k} \cdot S^r(\mathfrak{p}).$$

Proof #1. It is clearly sufficient to prove the first statement. Since $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^{\phi}$, $\mathfrak{p} = \mathfrak{a} \oplus \mathfrak{q}$, where $\mathfrak{q}$ is the span of $\{e - \theta e | e \in \mathfrak{g}^{\phi}, \phi \in \Sigma\}$. Now

$$S^r(\mathfrak{p}) = \coprod_{j=0}^{r} S^j(\mathfrak{q}) S^{r-j}(\mathfrak{a}).$$

It will be sufficient to prove by induction on j that the smallest $\mathfrak{k}$-invariant subspace T of $S^r(\mathfrak{p})$ containing $S^r(\mathfrak{a})$ also contains $S^j(\mathfrak{q})S^{r-j}(\mathfrak{a})$. This is clearly true for $j = 0$, so suppose it is true for $0, 1, \ldots, j$. We shall now prove it for $j + 1$. We assume that $j < r$.

Let $\phi \in \Sigma$, $e \in \mathfrak{g}^{\phi}$, $s \in S^j(\mathfrak{q})$ and $a \in \mathfrak{a}$. Then $e + \theta e \in \mathfrak{k}$, and

$$(e + \theta e) \cdot sa^{r-j} = ((e + \theta e) \cdot s)a^{r-j} - (r - j)\phi(a)s(e - \theta e)a^{r-(j+1)}$$

The left-hand side and the first term on the right are in T by the induction hypothesis, and so $\phi(a)s(e - \theta e)a^{r-(j+1)} \in T$. Let $Z = \{a \in \mathfrak{a} | \phi(a) \neq 0$ for all $\phi \in \Sigma\}$. Then Z is Zariski dense in $\mathfrak{a}$ since it is the subset of $\mathfrak{a}$ on which finitely many nonzero polynomial functions do not vanish; the fact that $S(\mathfrak{a}^*)$ is an integral domain implies easily that any such set is Zariski dense. Then for all $\phi \in \Sigma$, $e \in \mathfrak{g}^{\phi}$, $s \in S^j(\mathfrak{q})$ and $a \in Z$, $s(e - \theta e)a^{r-(j+1)} \in T$. But by Lemma 3.5(ii), $\{a^{r-(j+1)} | a \in Z\}$ spans $S^{r-(j+1)}(\mathfrak{a})$. Also, as ϕ, e and s vary, the terms $s(e - \theta e)$ span $S^{j+1}(\mathfrak{q})$. Thus we have shown that

$$S^{j+1}(\mathfrak{q})S^{r-(j+1)}(\mathfrak{a}) \subset T,$$

completing the induction. Q.E.D.

Proof #2 (McCollum). Use the first paragraph of Proof #1 and continue as follows:

Let $\phi \in \Sigma$ and choose $w \in \mathfrak{a}$ such that $\phi(w) = 1$. Let $\mathfrak{t} = \operatorname{Ker} \phi$, so that $\mathfrak{a} = kw \oplus \mathfrak{t}$. Now let $e \in \mathfrak{g}^{\phi}$, $s \in S^j(\mathfrak{q})$ and $t \in S^{r-j-i}(\mathfrak{t})$, where $1 \leq i \leq r - j$. Then $e + \theta e \in \mathfrak{k}$, and

$$(e + \theta e) \cdot sw^i t = ((e + \theta e) \cdot s)w^i t - is(e - \theta e)w^{i-1}t + sw^i((e + \theta e) \cdot t).$$

The left-hand side and the first term on the right are in T by the induction hypothesis, and the last term is zero because $t \in S(\mathfrak{t})$. Hence $s(e - \theta e)w^{i-1}t \in T$. But as i and t vary, the products $w^{i-1}t$ span $S^{r-(j+1)}(\mathfrak{a})$, so that $s(e - \theta e)S^{r-(j+1)}(\mathfrak{a}) \subset T$. Also, as ϕ, e and s vary, the products $s(e - \theta e)$ span $S^{j+1}(\mathfrak{q})$. Hence $S^{j+1}(\mathfrak{q})S^{r-(j+1)}(\mathfrak{a}) \subset T$, and this completes the induction. Q.E.D.

To complete the proof of the injectivity of F_*, let $f \in S(\mathfrak{p}^*)^{\mathfrak{k}}$ and assume $f | \mathfrak{a} = 0$. The homogeneous components of f are annihilated by $\mathfrak{k}$ since the decomposition

$$S(\mathfrak{p}^*) = \coprod_{j=0}^{\infty} S^r(\mathfrak{p}^*)$$

is a $\mathfrak{k}$-module decomposition. Also, the components of f vanish on $\mathfrak{a}$ because $\mathfrak{a}$ is closed under scalar multiplication. Hence we may assume $f \in S^r(\mathfrak{p}^*)^{\mathfrak{k}}$ for some $r \in \mathbf{Z}_+$. Consider the pairing $\{\cdot,\cdot\}$ between $S^r(\mathfrak{p}^*)$ and $S^r(\mathfrak{p})$. Now $\{f, a^r\} = 0$ for all $a \in \mathfrak{a}$, so that $\{f, S^r(\mathfrak{a})\} = 0$ by Lemma 3.5(ii). Also, for all $x \in \mathfrak{k}$ and $s \in S^r(\mathfrak{p})$, $\{f, x \cdot s\} = -\{x \cdot f, s\} = 0$, using Lemma 3.6. Thus $\{f, S^r(\mathfrak{p})\} = 0$ by Lemma 3.7, and $f = 0$ in view of the nonsingularity of $\{\cdot,\cdot\}$ (Lemma 3.5(i)). This proves the injectivity of F_*, and hence the theorem. Q.E.D.

Remark. In the Appendix, we shall give a vector-valued generalization of the injectivity of F_*.

In case $\dim \mathfrak{a} = 1$, we get more information. We now assume that $\dim \mathfrak{a} = 1$. The following result is clear:

Lemma 3.8. $S(\mathfrak{a}^*)^W$ *consists of the even polynomial functions on* $\mathfrak{a}$, *i.e., those polynomial functions* f *on* $\mathfrak{a}$ *such that* $f(a) = f(-a)$ *for all* $a \in \mathfrak{a}$. *If* f *is an arbitrary nonzero homogeneous quadratic polynomial function on* $\mathfrak{a}$, *then* f *generates* $S(\mathfrak{a}^*)^W$, *and* $S(\mathfrak{a}^*)^W = k[f]$ *is the polynomial algebra generated by* f.

The Killing form B of $\mathfrak{g}$ is nonsingular on $\mathfrak{a}$ (Lemma 2.2 or Lemma 3.2), and its restriction to $\mathfrak{p}$ is $\mathfrak{k}$-invariant. Thus the function b on $\mathfrak{p}$ defined by $p \mapsto B(p, p)$ is a homogeneous quadratic $\mathfrak{k}$-invariant polynomial function on $\mathfrak{p}$ whose restriction to $\mathfrak{a}$ is nonzero. Hence the last lemma implies:

Lemma 3.9. *The subalgebra* $k[b]$ *of* $S(\mathfrak{p}^*)^{\mathfrak{k}}$ *generated by* b *is the polynomial algebra generated by* b, *and* $F_*: k[b] \to S(\mathfrak{a}^*)^W$ *is an algebra isomorphism.*

In view of Theorem 3.1, we therefore have:

Theorem 3.10. *Suppose* $\dim \mathfrak{a} = 1$. *Then* $F_*: S(\mathfrak{p}^*)^{\mathfrak{k}} \to S(\mathfrak{a}^*)^W$ *is an algebra isomorphism, and* $S(\mathfrak{p}^*)^{\mathfrak{k}} = k[b]$; *here* $k[b]$ *is the polynomial algebra generated by* b.

Since the restriction of B to $\mathfrak{p}$ is nonsingular and $\mathfrak{k}$-invariant, B induces a $\mathfrak{k}$-module isomorphism from the contragredient $\mathfrak{k}$-module $\mathfrak{p}^*$ to $\mathfrak{p}$, and hence a $\mathfrak{k}$-module and algebra isomorphism from $S(\mathfrak{p}^*)$ to $S(\mathfrak{p})$. Let $b_0 \in S(\mathfrak{p})$ be the image of b under this isomorphism, so that b_0 is the canonical quadratic element of $S(\mathfrak{p})$ associated with the restriction of B to $\mathfrak{p}$, and b_0 is annihilated by $\mathfrak{k}$. Let $S(\mathfrak{p})^{\mathfrak{k}}$ denote the subalgebra of $\mathfrak{k}$-annihilated vectors of $S(\mathfrak{p})$. From Theorem 3.10, we have:

Corollary 3.11. *Suppose* $\dim \mathfrak{a} = 1$. *Then* $S(\mathfrak{p})^{\mathfrak{k}}$ *is generated by* b_0, *so that* $S(\mathfrak{p})^{\mathfrak{k}} = k[b_0]$, *and* $k[b_0]$ *is the polynomial algebra generated by* b_0.

4. The Harish-Chandra map p_*. Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the symmetric decomposition of a semisimple symmetric Lie algebra $(\mathfrak{g}, \theta)$ over a field k of characteristic zero, and let $\mathfrak{a}$ be a splitting Cartan subspace of $\mathfrak{p}$. Denote by $\Sigma \subset \mathfrak{a}^*$ the corresponding restricted root system and W the restricted Weyl group. Fix a positive system $\Sigma_+ \subset \Sigma$, and let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}$ be the corresponding Iwasawa decomposition. Define $\rho \in \mathfrak{a}^*$ by the condition

$$\rho(a) = \tfrac{1}{2} \operatorname{tr}(\operatorname{ad} a|\mathfrak{n})$$

for all $a \in \mathfrak{a}$, or equivalently,

$$\rho = \frac{1}{2} \sum_{\phi \in \Sigma_+} (\dim \mathfrak{g}^{\phi})\phi.$$

Let $\mathfrak{G}$ be the universal enveloping algebra of $\mathfrak{g}$, and let $\mathfrak{K}$, $\mathfrak{A}$ and $\mathfrak{N}$ be the universal enveloping algebras of $\mathfrak{k}$, $\mathfrak{a}$ and $\mathfrak{n}$, respectively, regarded as canonically embedded in $\mathfrak{G}$, by the Poincaré-Birkhoff-Witt theorem. The multiplication map in $\mathfrak{G}$ induces a linear isomorphism $\mathfrak{G} \simeq \mathfrak{K} \otimes \mathfrak{A} \otimes \mathfrak{N}$, by the same theorem. (In this section, $\otimes$ denotes tensor product over k.) Identifying $\mathfrak{G}$ with $\mathfrak{K} \otimes \mathfrak{A} \otimes \mathfrak{N}$, we have

$$\begin{aligned}\mathfrak{G} &= \mathfrak{K} \otimes \mathfrak{A} \otimes (k \cdot 1 \oplus \mathfrak{N}\mathfrak{n}) = \mathfrak{K} \otimes \mathfrak{A} \oplus \mathfrak{G}\mathfrak{n} \\ &= (k \cdot 1 \oplus \mathfrak{k}\mathfrak{K}) \otimes \mathfrak{A} \oplus \mathfrak{G}\mathfrak{n} = \mathfrak{A} \oplus \mathfrak{k}\mathfrak{K}\mathfrak{A} \oplus \mathfrak{G}\mathfrak{n}.\end{aligned}$$

Let $p: \mathfrak{G} \to \mathfrak{A}$ be the projection with respect to this decomposition. Since

$$\mathfrak{k}\mathfrak{G} = \mathfrak{k}\mathfrak{K}\mathfrak{A}\mathfrak{N} = \mathfrak{k}\mathfrak{K}\mathfrak{A}(k \cdot 1 + \mathfrak{N}\mathfrak{n}) \subset \mathfrak{k}\mathfrak{K}\mathfrak{A} + \mathfrak{G}\mathfrak{n},$$

we have

$$\mathfrak{G} = \mathfrak{A} \oplus (\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n}),$$

and p is also the projection to $\mathfrak{A}$ with respect to this decomposition.

Since $\mathfrak{a}$ is abelian, $\mathfrak{A}$ may be identified with the symmetric algebra $S(\mathfrak{a})$, and hence with the algebra of polynomial functions on $\mathfrak{a}^*$. Every affine automorphism T of $\mathfrak{a}^*$ gives rise to an algebra automorphism $T^\wedge$ of $\mathfrak{A} = S(\mathfrak{a})$, defined by $(T^\wedge f)(\lambda) = f(T^{-1}\lambda)$ for all $f \in \mathfrak{A}$, $\lambda \in \mathfrak{a}^*$. When T is translation by ρ, i.e., the map which takes $\lambda \in \mathfrak{a}^*$ to $\lambda + \rho$, we denote $T^\wedge$ by τ. Then $(\tau f)(\lambda) = f(\lambda - \rho)$ for all $f \in \mathfrak{A}$ and $\lambda \in \mathfrak{a}^*$, and τ may be characterized as the unique automorphism of $\mathfrak{A}$ which takes $a \in \mathfrak{a}$ to $a - \rho(a)$.

Now W is a group of linear automorphisms of $\mathfrak{a}^*$. For all $s \in W$, $s^\wedge$ is the unique automorphism of $\mathfrak{A}$ which acts on $\mathfrak{a}$ according to the contragredient of the action of s on $\mathfrak{a}^*$. Moreover, W acts as a group of automorphisms

of $\mathfrak{A}$ in this way. Let $\mathfrak{A}^W$ be the algebra of W-invariants in $\mathfrak{A}$.

Let $\mathfrak{G}^{\mathfrak{k}}$ denote the centralizer of $\mathfrak{k}$ in $\mathfrak{G}$, and let p_* be the map $(r \circ p)|\mathfrak{G}^{\mathfrak{k}}$, so that $p_* : \mathfrak{G}^{\mathfrak{k}} \to \mathfrak{A}$.

Theorem 4.1. *The map* $p_* : \mathfrak{G}^{\mathfrak{k}} \to \mathfrak{A}$ *is an algebra homomorphism with kernel* $\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G} = \mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}\mathfrak{k}$ *and image contained in* $\mathfrak{A}^W$.

Proof. The fact that p_* is a homomorphism is easy: Let $x, y \in \mathfrak{G}^{\mathfrak{k}}$, and write

$$x \equiv a \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})), \quad y \equiv b \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})),$$

where $a, b \in \mathfrak{A}$; then $a = p(x)$ and $b = p(y)$. We have

$$\begin{aligned} xy &\equiv xb \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})) \quad (\text{since } x \in \mathfrak{G}^{\mathfrak{k}}) \\ &\equiv ab \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})), \end{aligned}$$

since $[\mathfrak{a}, \mathfrak{n}] \subset \mathfrak{n}$. Hence

$$xy = p(x)p(y) \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})),$$

and so $p(xy) = p(x)p(y)$. Since r is a homomorphism, p_* is also a homomorphism.

Now $\mathrm{Ker}\, p_* = \mathrm{Ker}\, p|\mathfrak{G}^{\mathfrak{k}}$, and the fact that $\mathrm{Ker}\, p|\mathfrak{G}^{\mathfrak{k}} = \mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G} = \mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}\mathfrak{k}$ is proved in [1, Proposition 9.2.15]; the proof is essentially that of [5, Remark 4.6]. (Actually, the cited result deals with the projection to $\mathfrak{A}$ with respect to the decomposition $\mathfrak{G} = \mathfrak{A} \oplus (\mathfrak{G}\mathfrak{k} + \mathfrak{n}\mathfrak{G})$, but we get the desired result either by applying the transpose antiautomorphism of $\mathfrak{G}$ or by imitating the argument in [1] or [5] for the present map p.) The cited proof simplifies somewhat in the present special case. Also, in place of the argument in [5, Proof of Lemma 4.1] (see also [6] and [1, Lemma 9.2.7, part (b) of the proof]), which uses methods from the theory of Lie or algebraic groups, we can instead use the injectivity assertion in Theorem 3.1 above, whose proof of course does not involve Lie or algebraic groups. Incidentally, in the Appendix, we show how the injectivity argument in Theorem 3.1 can be used to give a proof of [1, Lemma 9.2.7(b)] in full generality, and even to generalize it, without algebraic groups. The cited proof of the equality $\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G} = \mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}\mathfrak{k}$ is independent of the assertion about the kernel of p_*. None of the above requires k to be algebraically closed.

We must now show that $p_*(\mathfrak{G}^{\mathfrak{k}}) \subset \mathfrak{A}^W$. We shall do this first when $\dim \mathfrak{a} = 1$, in which case we shall also show that the image of p_* is exactly $\mathfrak{A}^W$. We shall finally reduce the general case to this case.

Assume now that $\dim \mathfrak{a} = 1$. In proving that $p_*(\mathfrak{G}^{\mathfrak{k}}) = \mathfrak{A}^W$, we may, and do, also assume that k is algebraically closed. In fact, however, this assumption will only be used in proving Lemmas 4.2, 4.3 and 4.4.

Let $\lambda: S(\mathfrak{g}) \to \mathfrak{G}$ be the canonical linear isomorphism from the symmetric algebra of $\mathfrak{g}$ to the universal enveloping algebra, so that λ is defined by the formula

$$\lambda(g_1 \cdots g_n) = \frac{1}{n!} \sum_{\sigma} g_{\sigma(1)} \cdots g_{\sigma(n)}$$

for all $n \in \mathbf{Z}_+$ and $g_i \in \mathfrak{g}$; here the product on the left is taken in $S(\mathfrak{g})$, the products on the right are taken in $\mathfrak{G}$, and σ ranges through the group of permutations of $\{1, \ldots, n\}$ (see [1, §2.4]). For all $g \in \mathfrak{g}$ and $n \in \mathbf{Z}_+$, $\lambda(g^n) = g^n$, and this condition determines λ since the powers of elements of $\mathfrak{g}$ span $S(\mathfrak{g})$ (see Lemma 3.5(ii)). Moreover, λ is a $\mathfrak{g}$-module isomorphism with respect to the natural actions of $\mathfrak{g}$ on $S(\mathfrak{g})$ and $\mathfrak{G}$ as derivations (see [1, §2.4.10]).

Let B be the Killing form of $\mathfrak{g}$, so that B is nonsingular on $\mathfrak{p}$. Let b_0 be the canonical quadratic element of $S(\mathfrak{p})^{\mathfrak{k}}$ associated with the restriction of B to $\mathfrak{p}$, as at the end of §3, and set $u_0 = \lambda(b_0) \in \lambda(S(\mathfrak{p})^{\mathfrak{k}}) \subset \mathfrak{G}^{\mathfrak{k}}$. Our goal now is to compute $p_*(u_0)$.

As in §3, we define the nonsingular symmetric form B_θ on $\mathfrak{g}$ by $B_\theta(x, y) = -B(x, \theta y)$ for all $x, y \in \mathfrak{g}$. Then B_θ is nonsingular on $\mathfrak{a} \oplus \mathfrak{n}$, and $B_\theta(\mathfrak{a}, \mathfrak{n}) = 0$ (Lemma 3.2).

Lemma 4.2. *Define the linear map $f: \mathfrak{a} \oplus \mathfrak{n} \to \mathfrak{g}$ by the conditions $f = 1$ on $\mathfrak{a}$ and $f = 2^{-1/2}(1 - \theta)$ on $\mathfrak{n}$. Then $f(\mathfrak{a} \oplus \mathfrak{n}) \subset \mathfrak{p}$, and $f: \mathfrak{a} \oplus \mathfrak{n} \to \mathfrak{p}$ is a linear isomorphism which is an isometry from B_θ to B.*

Proof. Clearly, $f(\mathfrak{a} \oplus \mathfrak{n}) \subset \mathfrak{p}$. From the Iwasawa decomposition, it follows that $(1 - \theta): \mathfrak{a} \oplus \mathfrak{n} \to \mathfrak{p}$ is a linear isomorphism, and so $f: \mathfrak{a} \oplus \mathfrak{n} \to \mathfrak{p}$ is also a linear isomorphism, since $f = \frac{1}{2}(1 - \theta)$ on $\mathfrak{a}$ and $2^{-1/2}(1 - \theta)$ on $\mathfrak{n}$.

We must now show that for all $x, y \in \mathfrak{a} \oplus \mathfrak{n}$, $B(f(x), f(y)) = B_\theta(x, y)$. This is true if $x, y \in \mathfrak{a}$ since $B_\theta = B$ on $\mathfrak{a}$. Let $x \in \mathfrak{a}$, $y \in \mathfrak{n}$. Then $B_\theta(x, y) = 0$. But

$$B(f(x), f(y)) = 2^{-\frac{1}{2}} B(x, y - \theta y) = 2^{-\frac{1}{2}} B(x, y) - 2^{-\frac{1}{2}} B(x, \theta y) = 0,$$

so that the desired relation again holds. Finally, let $x, y \in \mathfrak{n}$. Then

$$B(f(x), f(y)) = \tfrac{1}{2} B(x - \theta x, y - \theta y) = -\tfrac{1}{2} B(x, \theta y) - \tfrac{1}{2} B(\theta x, y) = B_\theta(x, y).$$

The lemma follows from the bilinearity and symmetry of B_θ and B. Q.E.D.

Let $\alpha \in \Sigma_+$ be the (unique) simple restricted root, so that $\mathfrak{n} = \mathfrak{g}^\alpha \oplus \mathfrak{g}^{2\alpha}$, and $\mathfrak{g}^{2\alpha}$ may be zero. Since k is algebraically closed and B_θ is nonsingular on $\mathfrak{a}$, $\mathfrak{g}^\alpha$ and $\mathfrak{g}^{2\alpha}$ (Lemma 3.2), we may choose $e_1 \in \mathfrak{a}$ such that $B_\theta(e_1, e_1) = 1$ and B_θ-orthonormal bases $e_1^\alpha, \ldots, e_r^\alpha$ of $\mathfrak{g}^\alpha$ and $e_1^{2\alpha}, \ldots, e_s^{2\alpha}$ of $\mathfrak{g}^{2\alpha}$. The orthogonality of the decomposition $\mathfrak{a} \oplus \mathfrak{g}^\alpha \oplus \mathfrak{g}^{2\alpha}$ (Lemma 3.2) implies

that $e_1, e_1^\alpha, \dots, e_r^\alpha, e_1^{2\alpha}, \dots, e_s^{2\alpha}$ is a B_θ-orthonormal basis of $\mathfrak{a} \oplus \mathfrak{n}$. Hence by Lemma 4.2, $f(e_1), f(e_1^\alpha), \dots, f(e_r^\alpha), f(e_1^{2\alpha}), \dots, f(e_s^{2\alpha})$ is a B-orthonormal basis of $\mathfrak{p}$. This basis is $e_1, 2^{-1/2}(e_1^\alpha - \theta e_1^\alpha), \dots,$ $2^{-1/2}(e_1^{2\alpha} - \theta e_1^{2\alpha}), \dots$. But if $x_1, \dots, x_t$ is any B-orthonormal basis of $\mathfrak{p}$, $b_0 = x_1^2 + \cdots + x_t^2$ in $S(\mathfrak{p})$, and hence $u_0 = \lambda(b_0) = x_1^2 + \cdots + x_t^2$ in $\mathfrak{G}$. Thus

$$u_0 = e_1^2 + \tfrac{1}{2}(e_1^\alpha - \theta e_1^\alpha)^2 + \cdots + \tfrac{1}{2}(e_r^\alpha - \theta e_r^\alpha)^2$$

$$+ \tfrac{1}{2}(e_1^{2\alpha} - \theta e_1^{2\alpha})^2 + \cdots + \tfrac{1}{2}(e_s^{2\alpha} - \theta e_s^{2\alpha})^2.$$

To compute $p_*(u_0)$, first note that $p_*(e_1^2) = (\tau \circ p)(e_1^2) = \tau(e_1^2) = (e_1 - \rho(e_1))^2$. Now let $e = e_i^\alpha$ $(1 \le i \le r)$ or $e_j^{2\alpha}$ $(1 \le j \le s)$. Then

$$\begin{aligned} \tfrac{1}{2}(e - \theta e)^2 &= \tfrac{1}{2}(2e - (e + \theta e))^2 \\ &= \tfrac{1}{2}(4e^2 + (e + \theta e)^2 - 2e(e + \theta e) - 2(e + \theta e)e) \\ &\equiv -e(e + \theta e) \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})) \\ &\equiv -e\theta e \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})) \\ &\equiv -[e, \theta e] \ (\mathrm{mod}(\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n})). \end{aligned}$$

Recall from §3 that x_α is the unique element of $\mathfrak{a}$ such that $B(x_\alpha, x) = \alpha(x)$ for all $x \in \mathfrak{a}$ and that $x_{2\alpha} = 2x_\alpha$ (if $2\alpha \in \Sigma$). Since $B_\theta(e, e) = 1$, Lemma 3.3 implies that $[e, \theta e] = -x_\alpha$ if $e \in \mathfrak{g}^\alpha$ and $[e, \theta e] = -2x_\alpha$ if $e \in \mathfrak{g}^{2\alpha}$. Thus from the above computation,

$$p(u_0) = e_1^2 + (r + 2s)x_\alpha = e_1^2 + (\dim \mathfrak{g}^\alpha + 2 \dim \mathfrak{g}^{2\alpha})x_\alpha,$$

and hence

$$p_*(u_0) = (e_1 - \rho(e_1))^2 + (\dim \mathfrak{g}^\alpha + 2 \dim \mathfrak{g}^{2\alpha})(x_\alpha - \rho(x_\alpha)).$$

Let w be the unique element of $\mathfrak{a}$ such that $\alpha(w) = 1$. Then by the definition of ρ,

$$\rho(w) = \tfrac{1}{2}(\dim \mathfrak{g}^\alpha + 2 \dim \mathfrak{g}^{2\alpha}).$$

Let $w = ce_1$ and $x_\alpha = de_1$ $(c, d \in k)$. Then since $B(x_\alpha, w) = \alpha(w) = 1$ and $B(e_1, e_1) = B_\theta(e_1, e_1) = 1$, we have that $cd = 1$. Thus

$$\begin{aligned} p_*(u_0) &= (e_1 - \rho(e_1))^2 + 2\rho(w)(x_\alpha - \rho(x_\alpha)) \\ &= (e_1 - \rho(e_1))^2 + 2cd\rho(e_1)(e_1 - \rho(e_1)) \\ &= (e_1 - \rho(e_1))^2 + 2\rho(e_1)(e_1 - \rho(e_1)) \\ &= e_1^2 - \rho(e_1)^2. \end{aligned}$$

Summarizing, we have proved:

Lemma 4.3. *Assuming that* $\dim \mathfrak{a} = 1$ *and that* k *is algebraically closed, choose* $e_1 \in \mathfrak{a}$ *such that* $B(e_1, e_1) = 1$. *Then*

$$p_*(u_0) = e_1^2 - \rho(e_1)^2 = (e_1 + \rho(e_1))(e_1 - \rho(e_1)).$$

Since B is nonsingular on $\mathfrak{a}$, we can reformulate the last lemma as follows:

Lemma 4.4. *Assume* $\dim \mathfrak{a} = 1$, *and let* $e \in \mathfrak{a}$, $e \neq 0$. *Then* $B(e, e) \neq 0$, *and*

$$p_*(u_0) = (e^2 - \rho(e)^2)/B(e, e) = (e + \rho(e))(e - \rho(e))/B(e, e).$$

The point is that this holds even if k is not algebraically closed, and from now on we can drop the algebraic closure assumption.

Now clearly $e^2 - \rho(e)^2 \in \mathfrak{A}^W$, and since this element is quadratic (although not homogeneous), it generates $\mathfrak{A}^W$, and in fact $\mathfrak{A}^W = k[e^2 - \rho(e)^2]$ is the polynomial algebra generated by $e^2 - \rho(e)^2$. But $u_0 \in \mathfrak{G}^{\mathfrak{k}}$, and $p_*\colon \mathfrak{G}^{\mathfrak{k}} \to \mathfrak{A}$ is an algebra homomorphism. Hence we have:

Lemma 4.5. *The subalgebra* $k[u_0]$ *of* $\mathfrak{G}^{\mathfrak{k}}$ *generated by* u_0 *is isomorphic to the polynomial algebra generated by* u_0; $p_*(k[u_0]) \subset \mathfrak{A}^W$; *and* $p_*\colon k[u_0] \to \mathfrak{A}^W$ *is an algebra isomorphism. In particular,* $p_*(\mathfrak{G}^{\mathfrak{k}}) \supset \mathfrak{A}^W$.

To complete the proof that $p_*(\mathfrak{G}^{\mathfrak{k}}) = \mathfrak{A}^W$ when $\dim \mathfrak{a} = 1$, we need one last lemma:

Lemma 4.6. $\mathfrak{G}^{\mathfrak{k}} = (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) \oplus k[u_0]$.

Proof. Let $\mathfrak{G}_0 \subset \mathfrak{G}_1 \subset \mathfrak{G}_2 \subset \cdots$ be the standard filtration of $\mathfrak{G}$ and $S_0(\mathfrak{p}) \subset S_1(\mathfrak{p}) \subset S_2(\mathfrak{p}) \subset \cdots$ the standard filtration of $S(\mathfrak{p})$, so that for all $r \in \mathbf{Z}_+$, $S_r(\mathfrak{p}) = \coprod_{j=0}^{r} S^j(\mathfrak{p})$. The multiplication map in $\mathfrak{G}$ induces a linear isomorphism $\mathfrak{G} \simeq \mathfrak{K} \otimes \lambda(S(\mathfrak{p}))$, and $\mathfrak{G}_r \subset \mathfrak{K} \otimes \lambda(S_r(\mathfrak{p}))$ for all $r \in \mathbf{Z}_+$ (see [1, Proposition 2.4.15 and its proof]). Thus

$$\mathfrak{G}_r \subset (\mathfrak{k}\mathfrak{K} \oplus k \cdot 1) \otimes \lambda(S_r(\mathfrak{p})) \subset \mathfrak{k}\mathfrak{G} \oplus \lambda(S_r(\mathfrak{p})).$$

Since the decomposition on the right is a $\mathfrak{k}$-module decomposition,

$$\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}_r \subset (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) \oplus \lambda(S(\mathfrak{p})^{\mathfrak{k}} \cap S_r(\mathfrak{p})).$$

But $S(\mathfrak{p})^{\mathfrak{k}} = k[b_0]$ (Corollary 3.11), and so

$$(*) \qquad \mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}_r \subset (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) \oplus \lambda(k[b_0] \cap S_r(\mathfrak{p})).$$

Now the sum $(\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0]$ in the statement of the lemma is direct by Lemma 4.5, since $\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G} \subset \operatorname{Ker} p_*$. Hence to prove the lemma it is sufficient to show that $\mathfrak{G}^{\mathfrak{k}} \subset (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0]$. We shall show by induction on $r \in \mathbf{Z}_+$ that $\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}_r \subset (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0]$. This is trivial if $r = 0$. Assume it is true for r. To prove it for $r+1$, note that by (*) it is sufficient to show that

$$\lambda(k[b_0] \cap S_{r+1}(\mathfrak{p})) \subset (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0].$$

If r is even, we are done because $k[b_0] \cap S_{r+1}(\mathfrak{p}) = k[b_0] \cap S_r(\mathfrak{p})$, and the induction hypothesis implies the result. Suppose r is odd. In view of the induction hypothesis, it is sufficient to show that $\lambda(b_0^{(r+1)/2}) \in (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0]$. But

$$\lambda(b_0^{(r+1)/2}) \equiv \lambda(b_0^{(r-1)/2})\lambda(b_0) \pmod{\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{G}_r}.$$

(Indeed, for all $x \in S_m(\mathfrak{g})$, $y \in S_n(\mathfrak{g})$, we have $\lambda(xy) \equiv \lambda(x)\lambda(y) \pmod{\mathfrak{G}_{m+n-1}}$.) Again by the induction hypothesis, $\lambda(b_0^{(r-1)/2}) \subset (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0]$, so that

$$\lambda(b_0^{(r-1)/2})\lambda(b_0) = \lambda(b_0^{(r-1)/2})u_0 \in (\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}) + k[u_0].$$

A final application of the induction hypothesis proves the desired result. Q.E.D.

We now summarize our conclusions for the case $\dim \mathfrak{a} = 1$. From Lemmas 4.4, 4.5 and 4.6, we have:

Theorem 4.7. *Assume* $\dim \mathfrak{a} = 1$. *The homomorphism* $p_* : \mathfrak{G}^{\mathfrak{k}} \to \mathfrak{A}$ *has kernel* $\mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}$ *and image* $\mathfrak{A}^W$. *Let* b_0 *be the canonical quadratic element of* $S(\mathfrak{p})$ *associated with the restriction of the Killing form of* $\mathfrak{g}$ *to* $\mathfrak{p}$ *(see the end of* §3), *and let* $u_0 = \lambda(b_0)$, *so that* $u_0 \in \mathfrak{G}^{\mathfrak{k}}$. *Then the subalgebra* $k[u_0]$ *of* $\mathfrak{G}^{\mathfrak{k}}$ *generated by* u_0 *is isomorphic to the polynomial algebra generated by* u_0, *and* $p_* : k[u_0] \to \mathfrak{A}^W$ *is an algebra isomorphism. Moreover,*

$$p_*(u_0) = (e^2 - \rho(e)^2)/B(e, e),$$

where e *is an arbitrary nonzero element of* $\mathfrak{a}$.

Note that we did not have to refer to the general result on $\operatorname{Ker} p_*$ to show that $\operatorname{Ker} p_* = \mathfrak{G}^{\mathfrak{k}} \cap \mathfrak{k}\mathfrak{G}$ when $\dim \mathfrak{a} = 1$.

We must finally prove that $p_*(\mathfrak{G}^{\mathfrak{k}}) \subset \mathfrak{A}^W$ when $\dim \mathfrak{a}$ is arbitrary. We shall do this by applying Theorem 4.7 to certain semisimple subalgebras of $\mathfrak{g}$ associated with the simple restricted roots.

Assume then that $\dim \mathfrak{a}$ is arbitrary, and fix a simple restricted root α. Let $\mathfrak{m}$ be the centralizer of $\mathfrak{a}$ in $\mathfrak{k}$, and set

$$\mathfrak{g}_\alpha = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{j=\pm 1, \pm 2} \mathfrak{g}^{j\alpha} = \coprod_{j=-2}^{2} \mathfrak{g}^{j\alpha},$$

where $\mathfrak{g}^{2\alpha}$ and $\mathfrak{g}^{-2\alpha}$ might be zero. Then $\mathfrak{g}_\alpha$ is a subalgebra of $\mathfrak{g}$. Let Σ_α denote the set of positive restricted roots not proportional to α, and let

$$\mathfrak{n}^\alpha = \coprod_{\phi \in \Sigma_\alpha} \mathfrak{g}^\phi.$$

The simplicity of α implies that if $\beta \in \Sigma_\alpha$ and γ is either a positive restricted root or a restricted root proportional to α, then $\beta + \gamma \in \Sigma_\alpha$ if $\beta + \alpha$ is a restricted root. Hence $\mathfrak{n}^\alpha$ is a subalgebra of $\mathfrak{n}$, and $[\mathfrak{g}_\alpha, \mathfrak{n}^\alpha] \subset \mathfrak{n}^\alpha$. Also, setting $\mathfrak{n}_\alpha = \mathfrak{g}^\alpha \oplus \mathfrak{g}^{2\alpha}$, we have $\mathfrak{n} = \mathfrak{n}_\alpha \oplus \mathfrak{n}^\alpha$.

We claim that $\mathfrak{g}_\alpha$ is reductive in $\mathfrak{g}$. In fact, $\operatorname{Ker}\alpha$ is a subspace of $\mathfrak{a}$ and hence a subalgebra of $\mathfrak{g}$ reductive in $\mathfrak{g}$. But $\mathfrak{g}_\alpha$ is exactly the centralizer of $\operatorname{Ker}\alpha$ in $\mathfrak{g}$. The claim now follows from [1, Proposition 1.7.7].

Now $\mathfrak{g}_\alpha$ is stable under θ, so that $\mathfrak{g}_\alpha = \mathfrak{k}_\alpha \oplus \mathfrak{p}_\alpha$, where $\mathfrak{k}_\alpha = \mathfrak{g}_\alpha \cap \mathfrak{k}$ and $\mathfrak{p}_\alpha = \mathfrak{g}_\alpha \cap \mathfrak{p}$. Moreover, $\mathfrak{g}_\alpha$ is a reductive Lie algebra since it is reductive in $\mathfrak{g}$. Hence $\mathfrak{g}_1 = [\mathfrak{g}_\alpha, \mathfrak{g}_\alpha]$ is a semisimple Lie algebra and $\mathfrak{g}_\alpha = \mathfrak{g}_1 \oplus \mathfrak{c}$, where $\mathfrak{c}$ is the center of $\mathfrak{g}_\alpha$, and both $\mathfrak{g}_1$ and $\mathfrak{c}$ are stable under θ. Thus if we set $\theta_1 = \theta|\mathfrak{g}_1$, $\mathfrak{k}_1 = \mathfrak{g}_1 \cap \mathfrak{k}$, $\mathfrak{p}_1 = \mathfrak{g}_1 \cap \mathfrak{p}$, $\mathfrak{c}_+ = \mathfrak{c} \cap \mathfrak{k}$ and $\mathfrak{c}_- = \mathfrak{c} \cap \mathfrak{p}$, then $(\mathfrak{g}_1, \theta_1)$ is a semisimple symmetric Lie algebra with symmetric decomposition $\mathfrak{g}_1 = \mathfrak{k}_1 \oplus \mathfrak{p}_1$, and we also have $\mathfrak{c} = \mathfrak{c}_+ \oplus \mathfrak{c}_-$, $\mathfrak{k}_\alpha = \mathfrak{k}_1 \oplus \mathfrak{c}_+$ and $\mathfrak{p}_\alpha = \mathfrak{p}_1 \oplus \mathfrak{c}_-$.

Let $x_\alpha \in \mathfrak{a}$ be the unique element such that $B(x_\alpha, a) = \alpha(a)$ for all $a \in \mathfrak{a}$, where B is the Killing form of $\mathfrak{g}$. Then $\mathfrak{a} = kx_\alpha \oplus \operatorname{Ker}\alpha$.

Lemma 4.8. *We have*

$$\mathfrak{g}_1 = (\mathfrak{g}_1 \cap \mathfrak{m}) \oplus kx_\alpha \oplus \coprod_{j=\pm 1, \pm 2} \mathfrak{g}^{j\alpha}$$

and

$$\mathfrak{c} = (\mathfrak{c} \cap \mathfrak{m}) \oplus \operatorname{Ker}\alpha.$$

In particular, $\mathfrak{c}_+ = \mathfrak{c} \cap \mathfrak{m}$ *and* $\mathfrak{c}_- = \operatorname{Ker}\alpha$.

Proof. Clearly $\coprod \mathfrak{g}^{j\alpha} \subset \mathfrak{g}_1$, and so $\mathfrak{g}_1 = (\mathfrak{g}_1 \cap \mathfrak{m}) \oplus (\mathfrak{g}_1 \cap \mathfrak{a}) \oplus \coprod \mathfrak{g}^{j\alpha}$. To determine $\mathfrak{g}_1 \cap \mathfrak{a}$, recall that the symmetric bilinear form B_θ is nonsingular on $\mathfrak{g}^\alpha$ (see Lemma 3.2),and so we may choose $e \in \mathfrak{g}^\alpha$ such that $B_\theta(e, e) \neq 0$. But then Lemma 3.3 implies that $[e, \theta e]$ is a nonzero multiple of x_α. This shows that $kx_\alpha \subset \mathfrak{g}_1 \cap \mathfrak{a}$. On the other hand, $\operatorname{Ker}\alpha \subset \mathfrak{c}_-$, and $\mathfrak{c}_- \subset \mathfrak{a}$ because $\mathfrak{a}$ is its own centralizer in $\mathfrak{p}$. Since $\mathfrak{a} = kx_\alpha \oplus \operatorname{Ker}\alpha$ and $(\mathfrak{g}_1 \cap \mathfrak{a}) \cap \mathfrak{c}_- \subset \mathfrak{g}_1 \cap \mathfrak{c} = 0$, we must have $\mathfrak{g}_1 \cap \mathfrak{a} = kx_\alpha$ and $\mathfrak{c}_- = \operatorname{Ker}\alpha$. The rest of the lemma is clear. Q.E.D.

Lemma 4.9. *Let* $\mathfrak{a}_1 = kx_\alpha$. *Then* $\mathfrak{a}_1$ *is a splitting Cartan subspace of* $\mathfrak{p}_1$, *for the semisimple symmetric Lie algebra* $(\mathfrak{g}_1, \theta_1)$.

Proof. Clearly, $\mathfrak{a}_1$ is an abelian subspace of $\mathfrak{p}_1$ which is reductive in $\mathfrak{g}_1$. We must show that $\mathfrak{a}_1$ is its own centralizer in $\mathfrak{p}_1$. But from Lemma 4.8, the centralizer of $\mathfrak{a}_1$ in $\mathfrak{g}_1$ is $(\mathfrak{g}_1 \cap \mathfrak{m}) \oplus \mathfrak{a}_1$, and this implies the desired result. Q.E.D.

Let $\alpha_1 = \alpha|\mathfrak{a}_1$, so that $\alpha_1 \in \mathfrak{a}_1^*$. Then the restricted roots for $(\mathfrak{g}_1, \theta_1)$ with respect to $\mathfrak{a}_1$ are $\pm\alpha_1$ and possibly $\pm 2\alpha_1$ (depending on whether $\pm 2\alpha$ are roots for $\mathfrak{g}$). Choose α_1 (and possibly $2\alpha_1$) as the positive restricted roots, and let $\mathfrak{n}_1$ be the sum of the positive restricted root spaces in $\mathfrak{g}_1$. Then the corresponding Iwasawa decomposition of $\mathfrak{g}_1$ is $\mathfrak{g}_1 = \mathfrak{k}_1 \oplus \mathfrak{a}_1 \oplus \mathfrak{n}_1$. Moreover, $\mathfrak{n}_1 = \mathfrak{n}_\alpha$ as previously defined. Furthermore, the Iwasawa decomposition of $\mathfrak{g}_1$ is compatible with that of $\mathfrak{g}$, in the sense that $\mathfrak{k}_1 = \mathfrak{g}_1 \cap \mathfrak{k}$, $\mathfrak{a}_1 = \mathfrak{g}_1 \cap \mathfrak{a}$ and $\mathfrak{n}_1 = \mathfrak{g}_1 \cap \mathfrak{n}$.

Our goal now is to express the mapping $p: \mathcal{G} \to \mathcal{A}$ in a form which relates it to the corresponding mapping for $\mathfrak{g}_1$.

Let $\mathcal{G}_\alpha, \mathcal{K}_\alpha, \mathcal{N}_\alpha$ and $\mathcal{N}^\alpha$ denote the universal enveloping algebras of $\mathfrak{g}_\alpha$, $\mathfrak{k}_\alpha$, $\mathfrak{n}_\alpha$ and $\mathfrak{n}^\alpha$, respectively, regarded as canonically embedded in $\mathcal{G}$. Then regarding the following equalities as canonical linear isomorphisms, we have

$$\mathcal{G} = \mathcal{K} \otimes \mathcal{A} \otimes \mathcal{N} = \mathcal{K} \otimes \mathcal{A} \otimes \mathcal{N}_\alpha \otimes \mathcal{N}^\alpha = \mathcal{K} \otimes \mathcal{A} \otimes \mathcal{N}_\alpha \otimes (k \cdot 1 \oplus \mathcal{N}^\alpha \mathfrak{n}^\alpha)$$

$$= \mathcal{K} \otimes \mathcal{A} \otimes \mathcal{N}_\alpha \oplus \mathcal{G}\mathfrak{n}^\alpha.$$

Now let $\mathfrak{r}$ be an arbitrary linear complement of $\mathfrak{k}_\alpha$ in $\mathfrak{k}$, let $\lambda: S(\mathfrak{g}) \to \mathcal{G}$ denote the canonical linear isomorphism, and let $S_*(\mathfrak{r})$ denote the ideal $\amalg_{r=1}^\infty S^r(\mathfrak{r})$ of $S(\mathfrak{r})$. Then

$$\mathcal{K} = \lambda(S(\mathfrak{r})) \otimes \mathcal{K}_\alpha$$

(see [1, proof of Proposition 2.4.15])

$$= \lambda(k \cdot 1 \oplus S_*(\mathfrak{r})) \otimes \mathcal{K}_\alpha = (k \cdot 1 \oplus \lambda(S_*(\mathfrak{r}))) \otimes \mathcal{K}_\alpha$$

$$= \mathcal{K}_\alpha \oplus \lambda(S_*(\mathfrak{r})) \otimes \mathcal{K}_\alpha.$$

Hence

$$\mathcal{G} = (\mathcal{K}_\alpha \oplus \lambda(S_*(\mathfrak{r})) \otimes \mathcal{K}_\alpha) \otimes \mathcal{A} \otimes \mathcal{N}_\alpha \oplus \mathcal{G}\mathfrak{n}^\alpha$$

$$= \mathcal{K}_\alpha \otimes \mathcal{A} \otimes \mathcal{N}_\alpha \oplus \lambda(S_*(\mathfrak{r})) \otimes \mathcal{K}_\alpha \otimes \mathcal{A} \otimes \mathcal{N}_\alpha \oplus \mathcal{G}\mathfrak{n}^\alpha.$$

Since clearly $\mathfrak{g}_\alpha = \mathfrak{k}_\alpha \oplus \mathfrak{a} \oplus \mathfrak{n}_\alpha$, we have

$$\mathcal{G}_\alpha = \mathcal{K}_\alpha \otimes \mathcal{A} \otimes \mathcal{N}_\alpha,$$

and so

$$\mathfrak{G} = \mathfrak{G}_\alpha \oplus \lambda(S_*(\mathfrak{r})) \otimes \mathfrak{G}_\alpha \oplus \mathfrak{G}\mathfrak{n}^\alpha.$$

Let $q : \mathfrak{G} \to \mathfrak{G}_\alpha$ denote the projection map with respect to this decomposition. Then $\operatorname{Ker} q \subset \operatorname{Ker} p$, since $\lambda(S_*(\mathfrak{r})) \subset \mathfrak{k}\mathfrak{G}$ and $\mathfrak{G}\mathfrak{n}^\alpha \subset \mathfrak{G}\mathfrak{n}$, and so $p = p \circ q$.

Now $\mathfrak{g}_\alpha = \mathfrak{g}_1 \oplus \mathfrak{c}$, $\mathfrak{g}_1 = \mathfrak{k}_1 \oplus \mathfrak{a}_1 \oplus \mathfrak{n}_1$ and $\mathfrak{c} = \mathfrak{c}_+ \oplus \mathfrak{c}_-$. Letting $\mathfrak{G}_1$, $\mathfrak{A}_1$, $\mathfrak{C}$, $\mathfrak{C}_+$ and $\mathfrak{C}_-$ denote the universal enveloping algebras of $\mathfrak{g}_1$, $\mathfrak{a}_1$, $\mathfrak{c}$, $\mathfrak{c}_+$ and $\mathfrak{c}_-$, respectively, we have

$$\mathfrak{G}_1 = \mathfrak{A}_1 \oplus (\mathfrak{k}_1\mathfrak{G}_1 + \mathfrak{G}_1\mathfrak{n}_1)$$

and

$$\mathfrak{C} = \mathfrak{C}_+ \otimes \mathfrak{C}_- = (k \cdot 1 \oplus \mathfrak{c}_+\mathfrak{C}_+) \otimes \mathfrak{C}_- = \mathfrak{C}_- \oplus \mathfrak{c}_+\mathfrak{C}.$$

Let $p_1 : \mathfrak{G}_1 \to \mathfrak{A}_1$ and $p_2 : \mathfrak{C} \to \mathfrak{C}_-$ be the projections with respect to these decompositions, so that in particular, p_1 is the mapping for $\mathfrak{G}_1$ analogous to the mapping p for $\mathfrak{G}$. Now $\mathfrak{G}_\alpha = \mathfrak{G}_1 \otimes \mathfrak{C}$ and $\mathfrak{A} = \mathfrak{A}_1 \otimes \mathfrak{C}_-$, so we have a mapping $p_1 \otimes p_2 : \mathfrak{G}_\alpha \to \mathfrak{A}$.

Lemma 4.10. *The maps $p_1 \otimes p_2$ and $p|\mathfrak{G}_\alpha$ from $\mathfrak{G}_\alpha$ to $\mathfrak{A}$ are the same.*

Proof. Let $x \in \mathfrak{G}_1$, $y \in \mathfrak{C}$, and write $x \equiv a \pmod{\mathfrak{k}_1\mathfrak{G}_1 + \mathfrak{G}_1\mathfrak{n}_1}$ and $y \equiv b \pmod{\mathfrak{c}_+\mathfrak{C}}$, where $a \in \mathfrak{A}_1$ and $b \in \mathfrak{C}_-$. Then since $\mathfrak{C}$ centralizes $\mathfrak{G}_1$,

$$\begin{aligned} xy &\equiv ay \pmod{\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n}} \\ &\equiv ab \pmod{\mathfrak{k}\mathfrak{G} + \mathfrak{G}\mathfrak{n}}, \end{aligned}$$

and so $p(xy) = ab$. Q.E.D.

Hence we have:

Lemma 4.11. *The map $p : \mathfrak{G} \to \mathfrak{A}$ can be expressed in the form $p = (p_1 \otimes p_2) \circ q$.*

In order to complete our proof, we have to be more specific about the choice of the complement $\mathfrak{r}$ of $\mathfrak{k}_\alpha$ in $\mathfrak{k}$.

Lemma 4.12. *The subalgebra $\mathfrak{k}_\alpha$ is reductive in $\mathfrak{g}$.*

Proof. Since $\operatorname{Ker} \alpha$ is a subalgebra of $\mathfrak{g}$ reductive in $\mathfrak{g}$ and since $\mathfrak{g}_\alpha$ is the centralizer of $\operatorname{Ker} \alpha$ in $\mathfrak{g}$, [1, Proposition 1.7.7] implies that the restriction to $\mathfrak{g}_\alpha$ of the Killing form B of $\mathfrak{g}$ is nonsingular and that the semisimple and nilpotent components (with respect to $\mathfrak{g}$) of an element of $\mathfrak{g}_\alpha$ belong to $\mathfrak{g}_\alpha$. Now $B(\mathfrak{k}_\alpha, \mathfrak{p}_\alpha) = 0$, so that B is nonsingular on $\mathfrak{k}_\alpha$. Let $x \in \mathfrak{k}_\alpha$, and let x_s and x_n be the semisimple and nilpotent components of x, respectively. Then $x_s, x_n \in \mathfrak{g}_\alpha$ by the above. But $x = \theta x = \theta x_s + \theta x_n$, and since θx_s is semisimple, θx_n is nilpotent and $[\theta x_s, \theta x_n] = \theta[x_s, x_n] = 0$, we must

have $\theta x_s = x_s$ and $\theta x_n = x_n$. Hence $x_s \in \mathfrak{g}_\alpha \cap \mathfrak{k} = \mathfrak{k}_\alpha$ and $x_n \in \mathfrak{g}_\alpha \cap \mathfrak{k} = \mathfrak{k}_\alpha$. Thus $\mathfrak{k}_\alpha$ satisfies the conditions of [1, Proposition 1.7.6], and so $\mathfrak{k}_\alpha$ is reductive in $\mathfrak{g}$. Q.E.D.

By the lemma, $\mathfrak{k}_\alpha$ is reductive in $\mathfrak{k}$, and so we may choose $\mathfrak{r}$ above to be a $\mathfrak{k}_\alpha$-invariant complement of $\mathfrak{k}_\alpha$ in $\mathfrak{k}$. Then the three summands in the decomposition above which defines the projection q are all $\mathfrak{k}_\alpha$-stable (recall that $[\mathfrak{g}_\alpha, \mathfrak{n}^\alpha] \subset \mathfrak{n}^\alpha$), and so q is a $\mathfrak{k}_\alpha$-map. In particular, $q(\mathcal{G}^{\mathfrak{k}_\alpha}) = \mathcal{G}_\alpha^{\mathfrak{k}_\alpha}$, where superscript as usual denotes centralizer. Now since $\mathfrak{k}_\alpha = \mathfrak{k}_1 \oplus \mathfrak{c}_+$ and $\mathfrak{c}_+$ centralizes $\mathcal{G}_\alpha$, $\mathcal{G}_\alpha^{\mathfrak{k}_\alpha} = \mathcal{G}_\alpha^{\mathfrak{k}_1}$. But since $\mathcal{G}_\alpha = \mathcal{G}_1 \otimes \mathcal{C}$ and $\mathfrak{k}_1$ centralizes $\mathcal{C}$, $\mathcal{G}_\alpha^{\mathfrak{k}_1} = \mathcal{G}_1^{\mathfrak{k}_1} \otimes \mathcal{C}$. Hence from Lemma 4.11, we have

$$\begin{aligned} p(\mathcal{G}^{\mathfrak{k}_\alpha}) &= (p_1 \otimes p_2)(q(\mathcal{G}^{\mathfrak{k}_\alpha})) = (p_1 \otimes p_2)(\mathcal{G}_1^{\mathfrak{k}_1} \otimes \mathcal{C}) \\ &= p_1(\mathcal{G}_1^{\mathfrak{k}_1}) \otimes p_2(\mathcal{C}) = p_1(\mathcal{G}_1^{\mathfrak{k}_1}) \otimes \mathcal{C}_-. \end{aligned}$$

The conclusion is:

Lemma 4.13. *We have* $p(\mathcal{G}^{\mathfrak{k}_\alpha}) = p_1(\mathcal{G}_1^{\mathfrak{k}_1}) \otimes \mathcal{C}_-$.

We are now in a position to apply Theorem 4.7. Let

$$\rho_\alpha = \tfrac{1}{2}(\dim \mathfrak{g}^\alpha + 2 \dim \mathfrak{g}^{2\alpha})\alpha$$

and let $\rho'_\alpha = \rho_\alpha | \mathfrak{a}_1$. Then ρ'_α is half the sum of the positive restricted roots (with multiplicities counted) for $\mathfrak{g}_1$, and $\rho_\alpha | \mathfrak{c}_- = 0$. Let σ_α be the affine automorphism of $\mathfrak{a}^*$ which takes $\lambda \in \mathfrak{a}^*$ to $s_\alpha(\lambda + \rho_\alpha) - \rho_\alpha$ (where s_α is the Weyl reflection with respect to α), and let $\gamma = \hat{\sigma}_\alpha$ (in the sense of the beginning of this section), so that γ is an automorphism of $\mathcal{A}$. Also, let σ'_α be the affine automorphism of $\mathfrak{a}_1^*$ which takes $\lambda \in \mathfrak{a}_1^*$ to $-\lambda - 2\rho'_\alpha$, and let $\delta = (\sigma'_\alpha)^\wedge : \mathcal{A}_1 \to \mathcal{A}_1$. (Here the symbol $^\wedge$ is used with respect to $\mathfrak{a}_1$ instead of $\mathfrak{a}$.) Denote by $\mathcal{A}^\gamma$ and $\mathcal{A}_1^\delta$ the respective algebras of invariants. By Theorem 4.7, $p_1(\mathcal{G}_1^{\mathfrak{k}_1}) = \tau_1^{-1}(\mathcal{A}_1^{W_1})$, where W_1 denotes the (two-element) Weyl group of $\mathfrak{g}_1$, and $\tau_1 = T^\wedge : \mathcal{A}_1 \to \mathcal{A}_1$, where $T : \mathfrak{a}_1^* \to \mathfrak{a}_1^*$ is translation by ρ'_α. But $\tau_1^{-1}(\mathcal{A}_1^{W_1})$ is exactly $\mathcal{A}_1^\delta$, so that $p_1(\mathcal{G}_1^{\mathfrak{k}_1}) = \mathcal{A}_1^\delta$. On the other hand, we have:

Lemma 4.14. *If we identify* $\mathcal{A}$ *with* $\mathcal{A}_1 \otimes \mathcal{C}_-$, *the automorphism* γ *of* $\mathcal{A}$ *equals* $\delta \otimes 1$.

Proof. Let s'_α be the linear automorphism of $\mathfrak{a}$ which is the transpose of s_α. Then s'_α is -1 on $\mathfrak{a}_1$ and 1 on $\mathfrak{c}_-$. Now γ is the automorphism of $\mathcal{A}$ determined by the condition $\gamma(a) = s'_\alpha(a) - 2\rho_\alpha(a)$ for all $a \in \mathfrak{a}$. Also, δ is the automorphism of $\mathcal{A}_1$ determined by the condition $\delta(a_1) = -a_1 - 2\rho'_\alpha(a_1) = -a_1 - 2\rho_\alpha(a_1)$ for all $a_1 \in \mathfrak{a}_1$. Let $a_1 \in \mathfrak{a}_1$ and $c \in \mathfrak{c}_-$. It is sufficient to show that $\gamma(a_1 + c) = \delta \otimes 1(a_1 \otimes 1 + 1 \otimes c)$ (regarding $\mathcal{A}$ and $\mathcal{A}_1 \otimes \mathcal{C}$ as

identified). But

$$\gamma(a_1 + c) = s'_\alpha(a_1 + c) - 2\rho_\alpha(a_1 + c) = -a_1 + c - 2\rho_\alpha(a_1),$$

and

$$\delta \otimes 1(a_1 \otimes 1 + 1 \otimes c) = \delta(a_1) \otimes 1 + 1 \otimes c$$

$$= (-a_1 - 2\rho_\alpha(a_1)) \otimes 1 + 1 \otimes c.$$

Since these two elements identify with each other, the lemma is proved. Q.E.D.

In view of the lemma, $\mathfrak{A}^\gamma = \mathfrak{A}_1^\delta \otimes \mathcal{C}_-$, and so we can conclude from Lemma 4.13 and the discussion preceding Lemma 4.14:

Lemma 4.15. *We have* $p(\mathfrak{G}^{\mathfrak{k}_\alpha}) = \mathfrak{A}^\gamma$; *here* $\mathfrak{A}^\gamma$ *denotes the algebra of invariants in* $\mathfrak{A}$ *under the automorphism* $\gamma = \hat{\sigma_\alpha}$, *where* σ_α *is the affine automorphism of* $\mathfrak{a}^*$ *which takes* $\lambda \in \mathfrak{a}^*$ *to* $s_\alpha(\lambda + \rho_\alpha) - \rho_\alpha$.

We need one final lemma. Recall that $\rho = \frac{1}{2} \Sigma_{\phi \in \Sigma_+}(\dim \mathfrak{g}^\phi)\phi$.

Lemma 4.16. *We have* $s_\alpha(\rho) - \rho = s_\alpha(\rho_\alpha) - \rho_\alpha$.

Proof. Let Σ_α denote the set of positive restricted roots not proportional to α. Then the simplicity of α implies that $s_\alpha \Sigma_\alpha = \Sigma_\alpha$. On the other hand, for all $\beta \in \Sigma$, $s_\alpha\beta \in \Sigma$ and in fact $\dim \mathfrak{g}^\beta = \dim \mathfrak{g}^{s_\alpha\beta}$ (see Lemma 2.5). Thus

$$s_\alpha(\rho) = -\frac{1}{2}(\dim \mathfrak{g}^\alpha + 2 \dim \mathfrak{g}^{2\alpha})\alpha + \frac{1}{2} \sum_{\phi \in \Sigma_\alpha} (\dim \mathfrak{g}^\phi)\phi,$$

and so

$$s_\alpha(\rho) - \rho = -(\dim \mathfrak{g}^\alpha + 2 \dim \mathfrak{g}^{2\alpha})\alpha = -2\rho_\alpha = s_\alpha(\rho_\alpha) - \rho_\alpha. \qquad \text{Q.E.D.}$$

By the last two lemmas, $p(\mathfrak{G}^{\mathfrak{k}_\alpha}) = \mathfrak{A}^\gamma$, where $\gamma = \hat{\sigma_\alpha}$ and σ_α is the affine automorphism of $\mathfrak{a}^*$ which takes $\lambda \in \mathfrak{a}^*$ to $s_\alpha(\lambda + \rho) - \rho$. Recall that $\tau = T\hat{}$: $\mathfrak{A} \to \mathfrak{A}$ where T: $\mathfrak{a}^* \to \mathfrak{a}^*$ is translation by ρ. Then $\tau(\mathfrak{A}^\gamma)$ is the algebra of invariants for $\hat{s_\alpha}$. Denoting this algebra by $\mathfrak{A}^{s_\alpha}$, we now have the following conclusion:

Theorem 4.17. *In the notation of the beginning of this section, let* $\alpha \in \Sigma$ *be an arbitrary simple restricted root, define the subalgebra*

$$\mathfrak{g}_\alpha = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{j=\pm 1, \pm 2} \mathfrak{g}^{j\alpha} = \coprod_{j=-2}^{2} \mathfrak{g}^{j\alpha},$$

where $\mathfrak{g}^{2\alpha}$ *and* $\mathfrak{g}^{-2\alpha}$ *might be zero, and let* $\mathfrak{k}_\alpha = \mathfrak{g}_\alpha \cap \mathfrak{k}$. *Denote by* $\mathfrak{G}^{\mathfrak{k}_\alpha}$ *the centralizer of* $\mathfrak{k}_\alpha$ *in* $\mathfrak{G}$, *and by* $\mathfrak{A}^{s_\alpha}$ *the subalgebra of* $\mathfrak{A}$ *consisting of the*

elements invariant under the natural action of the Weyl reflection s_α *on* $\mathfrak{A}$. *Then* $(r \circ p)(\mathfrak{G}^{\mathfrak{k}_\alpha}) = \mathfrak{A}^{s_\alpha}$. *In particular,* $p_*(\mathfrak{G}^{\mathfrak{k}}) \subset \mathfrak{A}^{s_\alpha}$.

Since W is generated by the simple reflections s_α, we can conclude that $p_*(\mathfrak{G}^{\mathfrak{k}}) \subset \mathfrak{A}^W$, and Theorem 4.1 is proved. Q.E.D.

5. **Appendix.** Here we shall give a vector-valued generalization of the injectivity of the map F_* (see §3). In a class of important cases (of the module V, in the notation below), this generalization is already known (see [5, Lemma 4.1], [6] and [1, Lemma 9.2.7, part (b) of the proof]), but in addition to being more general, the present proof is elementary in that it does not require theory of Lie or algebraic groups. The argument presented here is G. McCollum's simplification of our original proof.

Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the symmetric decomposition of a semisimple symmetric Lie algebra over a field k of characteristic zero and let $\mathfrak{a}$ be a Cartan subspace of $\mathfrak{p}$. Then $\mathfrak{k}$ acts naturally on $\mathfrak{p}$, and thus on $\mathfrak{p}^*$ and on $S(\mathfrak{p}^*)$. Let V be an arbitrary (possibly infinite-dimensional) $\mathfrak{k}$-module. Then since $S(\mathfrak{p}^*)$ is naturally identified with the algebra of polynomial functions on $\mathfrak{p}$, the tensor product $\mathfrak{k}$-module $S(\mathfrak{p}^*) \otimes V$ may be identified with a space of V-valued functions on $\mathfrak{p}$. (In this section, $\otimes$ denotes tensor product over k.) Similarly, $S(\mathfrak{a}^*) \otimes V$ may be identified with a space of V-valued functions on $\mathfrak{a}$. Let $(S(\mathfrak{p}^*) \otimes V)^{\mathfrak{k}}$ be the space of $\mathfrak{k}$-annihilated vectors in $S(\mathfrak{p}^*) \otimes V$, and let $F_*^V : (S(\mathfrak{p}^*) \otimes V)^{\mathfrak{k}} \to S(\mathfrak{a}^*) \otimes V$ denote the natural restriction map.

Theorem 5.1. F_*^V *is injective.*

Proof. Let $f_0 \in (S(\mathfrak{p}^*) \otimes V)^{\mathfrak{k}}$ and suppose $F_*^V(f_0) = 0$. The homogeneous components of f_0 with respect to the decomposition

$$S(\mathfrak{p}^*) \otimes V = \coprod_{r=0}^{\infty} S^r(\mathfrak{p}^*) \otimes V$$

are annihilated by $\mathfrak{k}$, since the terms in this decomposition are $\mathfrak{k}$-stable. The components of f_0 also vanish on $\mathfrak{a}$; this follows from the fact that $\mathfrak{a}$ is stable under scalar multiplication. Hence it is sufficient to prove that if f_0 is a $\mathfrak{k}$-annihilated element of $S^r(\mathfrak{p}^*) \otimes V$ for some $r \in \mathbf{Z}_+$ and if the restriction of f_0 to $\mathfrak{a}$ is zero, then $f_0 = 0$.

Recall from §3 the pairing $\{\cdot,\cdot\}$ between $S^r(\mathfrak{p}^*)$ and $S^r(\mathfrak{p})$. Define a bilinear map $\omega : (S^r(\mathfrak{p}^*) \otimes V) \times S^r(\mathfrak{p}) \to V$ by the condition $g \otimes v, s \mapsto \{g, s\}v$ for all $g \in S^r(\mathfrak{p}^*)$, $v \in V$ and $s \in S^r(\mathfrak{p})$. In view of Lemma 3.6, ω is a $\mathfrak{k}$-map in the sense that $\omega(x \cdot f, s) + \omega(f, x \cdot s) = x \cdot \omega(f, s)$ for all $x \in \mathfrak{k}$, $f \in S^r(\mathfrak{p}^*) \otimes V$ and $s \in S^r(\mathfrak{p})$. Also, for all $f \in S^r(\mathfrak{p}^*) \otimes V$, $\omega(f, S^r(\mathfrak{p})) = 0$ implies $f = 0$. Indeed, let $\{v_i\}$ be a basis of V, and write $f = \Sigma_i\, g_i \otimes v_i$, where $g_i \in S^r(\mathfrak{p}^*)$.

Then

$$0 = \sum_i \omega(g_i \otimes v_i, S^r(\mathfrak{p})) = \sum_i \{g_i, S^r(\mathfrak{p})\} v_i$$

so that $\{g_i, S^r(\mathfrak{p})\} = 0$ for each i. By Lemma 3.5(i), each $g_i = 0$, and so $f = 0$.

Now let $T = \{t \in S^r(\mathfrak{p}) | \omega(f_0, t) = 0\}$. Then T is a $\mathfrak{k}$-submodule of $S^r(\mathfrak{p})$. In fact, if $t \in T$ and $x \in \mathfrak{k}$, then $\omega(f_0, x \cdot t) = x \cdot \omega(f_0, t) - \omega(x \cdot f_0, t) = 0$ since $t \in T$ and $x \cdot f_0 = 0$. But $S^r(\mathfrak{a}) \subset T$. Indeed, let $a \in \mathfrak{a}$, and write $f_0 = \Sigma_i\, g_i \otimes v_i$ for some $g_i \in S^r(\mathfrak{p}^*)$ and $v_i \in V$. Then

$$\omega(f_0, a^r) = \sum_i \{g_i, a^r\} v_i = \sum_i r!\, g_i(a) v_i = r!\, f_0(a) = 0$$

by hypothesis, and the fact that $S^r(\mathfrak{a}) \subset T$ follows from Lemma 3.5(ii). Thus T is a $\mathfrak{k}$-submodule of $S^r(\mathfrak{p})$ containing $S^r(\mathfrak{a})$, so that $T = S^r(\mathfrak{p})$ by Lemma 3.7. (Note that the field extension technique shows that Lemma 3.7 holds even when $\mathfrak{a}$ is not a splitting Cartan subspace.) That is, $\omega(f_0, S^r(\mathfrak{p})) = 0$, and so $f_0 = 0$ by the last paragraph. Q. E. D.

BIBLIOGRAPHY

1. J. Dixmier, *Algèbres enveloppantes*, Gauthier-Villars, Paris, 1974.

2. Harish-Chandra, (a) *Representations of semisimple Lie groups*. II, Trans. Amer. Math. Soc. 76 (1954), 26–65. MR 15, 398.

(b) *Spherical functions on a semisimple Lie group*. I, Amer. J. Math. 80 (1958), 241–310. MR 20 #925.

3. S. Helgason, *A duality for symmetric spaces with applications to group representations*, Advances in Math. 5 (1970), 1–154. MR 44 #8587.

4. B. Kostant, *On the existence and irreducibility of certain series of representations*, Publ. 1971 Summer School in Math., edited by I. M. Gel'fand, Bolyai-János Math. Soc., Budapest (to appear).

5. J. Lepowsky, *Algebraic results on representations of semisimple Lie groups*, Trans. Amer. Math. Soc. 176 (1973), 1–44.

6. C. Rader, *Spherical functions on semisimple Lie groups*, Thesis and unpublished supplements, University of Washington, 1971.

7. Séminaire Sophus Lie École Norm. Sup. 1954/55, *Théorie des algèbres de Lie. Topologies des groupes de Lie*, Secretariat mathématique, Paris, 1955. MR 17, 384.

8. G. Warner, *Harmonic analysis on semi-simple Lie groups*. I, Springer-Verlag, New York, 1972.

DEPARTMENT OF MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, CONNECTICUT 06520

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

CONICAL VECTORS IN INDUCED MODULES

BY

J. LEPOWSKY(1)

ABSTRACT. Let $\mathfrak{g}$ be a real semisimple Lie algebra with Iwasawa decomposition $\mathfrak{g}=\mathfrak{k}\oplus\mathfrak{a}\oplus\mathfrak{n}$, and let $\mathfrak{m}$ be the centralizer of $\mathfrak{a}$ in $\mathfrak{k}$. A conical vector in a $\mathfrak{g}$-module is defined to be a nonzero $\mathfrak{m}\oplus\mathfrak{n}$-invariant vector. The $\mathfrak{g}$-modules which are algebraically induced from one-dimensional $(\mathfrak{m}\oplus\mathfrak{a}\oplus\mathfrak{n})$-modules on which the action of $\mathfrak{m}$ is trivial have "canonical generators" which are conical vectors. In this paper, all the conical vectors in these $\mathfrak{g}$-modules are found, in the special case $\dim\mathfrak{a}=1$. The conical vectors have interesting expressions as polynomials in two variables which factor into linear or quadratic factors. Because it is too difficult to determine the conical vectors by direct computation, metamathematical "transfer principles" are proved, to transfer theorems about conical vectors from one Lie algebra to another; this reduces the problem to a special case which can be solved. The whole study is carried out for semisimple symmetric Lie algebras with splitting Cartan subspaces, over arbitrary fields of characteristic zero. An exposition of the Kostant-Mostow double transitivity theorem is included.

1. Introduction. The theory of Verma modules, as developed by D.-N. Verma [10(a), (b)] and by I. N. Bernšteĭn, I. M. Gel′fand and S. I. Gel′fand [1(a), (b)], is becoming increasingly important. Let $\mathfrak{g}$ be a complex semisimple Lie algebra and $\mathfrak{b}$ a Borel subalgebra of $\mathfrak{g}$. The associated *Verma modules* are the $\mathfrak{g}$-modules induced, in the algebraic sense, by the one-dimensional $\mathfrak{b}$-modules (see [2, Chapter 7]). As we shall see in this introduction, a corresponding theory of $\mathfrak{g}$-modules induced from more general parabolic subalgebras of $\mathfrak{g}$ should also be developed, and the purpose of this paper is to begin such a study.

Here is our main reason for interest in this problem: Let $G = KAN$ be an Iwasawa decomposition of a real semisimple Lie group with finite center, and

Received by the editors March 14, 1974 and, in revised form, July 15, 1974.
AMS (MOS) subject classifications (1970). Primary 17B10; Secondary 16A64, 17B20, 17B35, 22E45.
Key words and phrases. Conical vectors, highest weight vectors, induced modules, Verma modules, real semisimple Lie algebras, real rank one, semisimple symmetric Lie algebras, splitting Cartan subspaces, restricted roots, restricted weight vectors, restricted Weyl group, universal enveloping algebra, double transitivity theorem, polynomial invariants.
(1) Partially supported by NSF GP-33893.

$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}$ the corresponding decomposition of the complexified Lie algebra of G. Let M be the centralizer of A in K, and $\mathfrak{m}$ its complexified Lie algebra. The infinitesimal nonunitary principal series of G is the family of $\mathfrak{g}$-modules obtained by taking the K-finite subspaces of the nonunitary principal series representations—those Hilbert space representations of G induced from the finite-dimensional irreducible representations of MAN (see for example [7(a)]). This family of $\mathfrak{g}$-modules is of great importance because every irreducible $\mathfrak{g}$-module which splits into a direct sum of finite-dimensional irreducible $\mathfrak{k}$-modules exponentiating to K-modules is a subquotient of an infinitesimal nonunitary principal series module (see [4], [7(a)], [9] and [2, Chapter 9]). But roughly speaking, the infinitesimal nonunitary principal series modules may be identified with certain "large" subspaces of the contragredient $\mathfrak{g}$-modules to $\mathfrak{g}$-modules *algebraically* induced by finite-dimensional irreducible modules of the parabolic subalgebra $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ of $\mathfrak{g}$ (cf. [2, §§9.3.1, 9.7.10]). Other important families of induced representations of G are similarly related to $\mathfrak{g}$-modules algebraically induced from parabolic subalgebras of $\mathfrak{g}$.

In a sense, the algebraically induced modules may be thought of as modules of distributions supported at the identity element of G, and their duals—algebraically "produced" modules—as modules of formal power series at the identity element of G. The K-finite elements of the produced modules (the K-finite formal power series) then correspond to analytic functions on G which are also the K-finite elements of the Hilbert space induced representations.

The Verma modules that can be embedded in a given Verma module are completely known ([10] and [1(a)]; see also [2, Théorème 7.6.23]). Suppose one could correspondingly determine the $\mathfrak{g}$-module maps between pairs of $\mathfrak{g}$-modules algebraically induced from $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$. Looking at the dual maps between the K-finite subspaces of the contragredient modules, one would have intertwining operators between nonunitary principal series G-modules, and these intertwining operators, which might be Kunze-Stein integral operators, would now be given by *differential* formulas. Furthermore, since an algebraically induced module is generated by a "highest weight vector" ($\mathfrak{n}$-invariant vector), the $\mathfrak{g}$-maps from one of the algebraically induced modules to another are closely related to the highest weight vectors in the target module. These give rise to highest weight vectors in the dual of the K-finite subspace of the Hilbert space induced G-module, and therefore are intimately connected with S. Helgason's conical distributions [5(a), (b)].(2) The submodule structure of

(2) See also M. Hu's thesis [12], whose results on conical distributions are related to our results on conical vectors.

the algebraically induced $\mathfrak{g}$-modules must also shed light on the subquotient structure of the nonunitary principal series modules (see M. Duflo [3] and [2, §9.6] for the case of complex G, using Verma modules), but examples show that the relation will be subtle. For instance, irreducibility of the algebraically induced module is not equivalent to irreducibility of the related contragredient nonunitary principal series module. On the other hand, the subquotient structure of the nonunitary principal series is notoriously complicated, but the structure of the algebraically induced modules already appears to be more regular and perhaps more fundamental. For example, the inclusion relations among the Verma submodules of certain Verma modules recover the inclusion relations among the closures of the Bruhat cells for complex semisimple Lie groups (see [10]), and it is likely that this situation will generalize to real semisimple Lie groups, using the modules algebraically induced from $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$.

Now that we want to find the highest weight vectors in a given $\mathfrak{g}$-module X algebraically induced from a finite-dimensional irreducible $(\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n})$-module, how do we do it? The following seemed at first like a good starting point: Let $\mathfrak{t}$ be a Cartan subalgebra of $\mathfrak{m}$, so that $\mathfrak{h} = \mathfrak{t} \oplus \mathfrak{a}$ is a Cartan subalgebra of $\mathfrak{g}$. Let $\mathfrak{b}$ be a Borel subalgebra of $\mathfrak{g}$ containing $\mathfrak{h}$ and $\mathfrak{n}$. Then it is easy to see that X is a $\mathfrak{g}$-module quotient of a certain Verma module V induced from $\mathfrak{b}$ (cf. [2, Lemma 9.3.2]). Hence one can try to use the well-developed theory of highest weight vectors in Verma modules to study highest weight vectors in X. Unfortunately, however, highest weight vectors in V can vanish when one passes to the quotient X, even in simple examples. Moreover, it turns out that there are, in general, highest weight vectors in X which do not come from highest weight vectors in V. This subtlety, which made the problem much more difficult than we expected it to be, forced us to work in a relatively special case and to develop new tools to handle even this case.

Now we shall describe our main results, and then we shall say what is interesting about our methods.

By analogy with Helgason's conical distributions, we call a nonzero vector in a $\mathfrak{g}$-module (or more generally, in an $\mathfrak{m} \oplus \mathfrak{n}$-module) *conical* if it is $\mathfrak{m} \oplus \mathfrak{n}$-invariant. The space of conical vectors, together with 0, is called the *conical space* of the module. Let $\mathfrak{G}$ be the universal enveloping algebra of $\mathfrak{g}$ and $\mathfrak{P} \subset \mathfrak{G}$ the universal enveloping algebra of $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$. Define $\rho \in \mathfrak{a}^*$ (* denotes dual) by the condition $\rho(a) = \frac{1}{2}\mathrm{tr}(\mathrm{ad}\ a|\mathfrak{n})$ for all $a \in \mathfrak{a}$, so that ρ is half the sum of the positive restricted roots with multiplicities counted. For all $\nu \in \mathfrak{a}^*$, the linear functional on $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ which is zero on $\mathfrak{m} \oplus \mathfrak{n}$ and

$\nu - \rho$ on $\mathfrak{a}$ defines a one-dimensional representation of $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$. Regarding $\mathbf{C}$ as the associated one-dimensional $\mathcal{P}$-module, and $\mathcal{G}$ as a right $\mathcal{P}$-module by right multiplication, we can form the $\mathcal{G}$-module $X^\nu = \mathcal{G} \otimes_{\mathcal{P}} \mathbf{C}$. This is a "twisted induced module" in the sense of [2, §5.2]. The vector $x_0 = 1 \otimes 1 \in X^\nu$ is a conical vector which generates X^ν, and is called the *canonical generator* of X^ν. Let $\mathfrak{n}^- \subset \mathfrak{g}$ be the sum of the negative restricted root spaces of $\mathfrak{g}$ with respect to $\mathfrak{a}$, and $\mathcal{N}^- \subset \mathcal{G}$ its universal enveloping algebra. Then $X^\nu = \mathcal{N}^- \cdot x_0$.

We are aiming for a description of the conical vectors in X^ν in case G has real rank 1, i.e., $\dim \mathfrak{a} = 1$. Assume this, and let $\alpha \in \mathfrak{a}^*$ be the unique simple restricted root. Then $\mathfrak{n}^-$ is the direct sum of the restricted root spaces $\mathfrak{g}^{-\alpha}$ and $\mathfrak{g}^{-2\alpha}$; here $\mathfrak{g}^{-2\alpha}$ may be zero. There are natural M-invariant nonsingular symmetric bilinear forms on $\mathfrak{g}^{-\alpha}$ and $\mathfrak{g}^{-2\alpha}$. Let $q_{-\alpha} \in \mathcal{N}^-$ and $q_{-2\alpha} \in \mathcal{N}^-$ be the sums of the squares of orthonormal bases of $\mathfrak{g}^{-\alpha}$ and $\mathfrak{g}^{-2\alpha}$, respectively, so that $q_{-\alpha}$ and $q_{-2\alpha}$ are quadratic M-invariant elements of $\mathcal{N}^-$, and $q_{-2\alpha} = 0$ if $\mathfrak{g}^{-2\alpha} = 0$. Let $(\mathcal{N}^-)^M$ be the algebra of all M-invariants in $\mathcal{N}^-$. Then $(\mathcal{N}^-)^M$ is a polynomial algebra on either one or two generators, depending on whether $\mathfrak{g}^{-2\alpha} = 0$ or $\mathfrak{g}^{-2\alpha} \neq 0$ (see §5), and in the difficult case when $\dim \mathfrak{g}^{-2\alpha} > 1$, the two generators are $q_{-\alpha}$ and $q_{-2\alpha}$; this follows from the Kostant-Mostow double transitivity theorem (see §4) on M-orbits in $\mathfrak{n}^-$ (or more precisely, M-orbits in the intersection of $\mathfrak{n}^-$ with the real Lie algebra of G). With this as background, we now state our main results (see §10):

Theorem 1.1. *Assume* $\dim \mathfrak{a} = 1$ *and let* $\nu \in \mathfrak{a}^*$. *Then the conical space of* X^ν *is either one- or two-dimensional, according to whether* ν *is a positive integral multiple of* α (*of* $\frac{1}{2}\alpha$ *if* $\dim \mathfrak{g}^\alpha = 1$) *or not. If* ν *is not of this form, then the conical space of* X^ν *is spanned by the canonical generator* x_0 *of* X^ν. *Suppose* $\nu = l\alpha$, l *a positive integer.* (*If* $\dim \mathfrak{g}^\alpha = 1$, *take instead* $\nu = \frac{1}{2}l\alpha$.) *Then* $q_{-\alpha}$ *and* $q_{-2\alpha}$ *can be suitably renormalized* (*independently of* l) *so that the following is true: Suppose* $\dim \mathfrak{g}^\alpha > 1$. *Define* $\zeta_l \in \mathcal{N}^-$ *by the formula*

$$\zeta_l = \begin{cases} \displaystyle\prod_{j=1;\, j\ odd}^{l-1} (q_{-\alpha}^2 + j^2 q_{-2\alpha}), & l\ even, \\[2ex] \displaystyle q_{-\alpha} \prod_{j=2;\, j\ even}^{l-1} (q_{-\alpha}^2 + j^2 q_{-2\alpha}), & l\ odd. \end{cases}$$

If $\dim \mathfrak{g}^\alpha = 1$, *define* $\zeta_l = f^l \in \mathcal{N}^-$, *where* f *is a nonzero element of* $\mathfrak{g}^{-\alpha}$. *Then the conical space of* X^ν *has basis* $\{x_0, \zeta_l \cdot x_0\}$. *Moreover, the* $\mathfrak{g}$-*submodule of* X^ν *generated by* $\zeta_l \cdot x_0$ *is isomorphic to* $X^{-\nu}$.

Theorem 1.2. *Let* $\mu, \nu \in \mathfrak{a}^*$. *Then* $\dim \mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu) \leq 1$. *Moreover,* $\dim \mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu) = 1$ *if and only if either* $\mu = \nu$, *or else* $\mu = -\nu$ *and* ν *is a nonnegative integral multiple of* α (*of* $\frac{1}{2}\alpha$ *if* $\dim \mathfrak{g}^\alpha = 1$). *This is exactly the case in which* X^μ *is isomorphic to a* $\mathfrak{g}$*-submodule of* X^ν.

(The annoying exceptional case $\dim \mathfrak{g}^\alpha = 1$ in these two theorems is essentially the case $G = SL(2, \mathbf{R})$, and is trivial.)

Considering how rare it is for a polynomial in two variables to factor into linear or quadratic factors, the factored form of the ζ_l in Theorem 1.1 seems remarkable. We shall say more about this below.

It turns out that Theorem 1.2 follows easily from Theorem 1.1, so we shall explain what is involved in proving Theorem 1.1. First, it is easy to see that the space of $\mathfrak{m}$-invariants in X^ν is the space $(\mathfrak{N}^-)^{\mathfrak{m}} \cdot x_0$ (here $(\mathfrak{N}^-)^{\mathfrak{m}}$ is the space of $\mathfrak{m}$-invariants in $\mathfrak{N}^-$ and equals $(\mathfrak{N}^-)^M$). From the above, $(\mathfrak{N}^-)^{\mathfrak{m}}$ is a polynomial algebra in one or two generators. If $\mathfrak{g}^{2\alpha} = 0$, we have one generator, and Theorem 1.1 is not terribly hard in this case (see §6). Suppose now that $\dim \mathfrak{g}^{2\alpha} > 1$, so that $(\mathfrak{N}^-)^{\mathfrak{m}}$ is the polynomial algebra $\mathbf{C}[q_{-\alpha}, q_{-2\alpha}]$. The whole problem is to determine those polynomials p in two variables such that $p(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ is $\mathfrak{n}$-invariant. Clearly, this involves computing commutators of elements of $\mathfrak{n}$ with $q_{-\alpha}$ and $q_{-2\alpha}$, and also commutators of *these* commutators with $q_{-\alpha}$ and $q_{-2\alpha}$. We were able to compute the necessary commutators (see §§6, 7), but the resulting condition on the polynomial p is immensely complicated, and it is not feasible to analyze it directly (see the last remark in §8).

However, when attempting to unravel this condition on p for some special G's, we noticed that the computations, even though we could not do them for any one G, did not seem to depend on G. The key was then to prove *a priori* that the conical vectors would look the same for any one G (for which $\dim \mathfrak{g}^{2\alpha} > 1$) as for any other such G, and then to use possibly special methods to solve the problem for one "small" G. Specifically, we first proved what we call the "fundamental commutation relation in $\mathfrak{N}^-$": There is a nonzero constant $c \in \mathbf{C}$ such that $[[f, q_{-\alpha}], q_{-\alpha}] = cfq_{-2\alpha}$ for all $f \in \mathfrak{g}^{-\alpha}$ (see Theorem 7.4). This is called "fundamental" because of the next result: If f is chosen more carefully, then this relation and a trivial one ($[f, q_{-2\alpha}] = 0$) generate *all* relations which are linear in f in the associative subalgebra of $\mathfrak{N}^-$ generated by f, $q_{-\alpha}$ and $q_{-2\alpha}$ (see Theorem 8.1). This in turn implies the following metamathematical "transfer principle for $\mathfrak{N}^-$": If $a_1, \dots, a_r$, $b_1, \dots, b_r$ are complex polynomials in two variables, then the truth of any assertion of the form "$\sum_{i=1}^r a_i(q_{-\alpha}, q_{-2\alpha}) f b_i(q_{-\alpha}, q_{-2\alpha}) = 0$ in $\mathfrak{N}^-$" is in-

dependent of G (see Theorem 8.4). But the condition that $p(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ be conical in X^ν can be expressed in this form (see Lemma 8.5), where the a_i and b_i depend only on p and the complex number c such that $\nu = c\alpha$. Thus we could prove the "transfer principle for conical vectors", another metatheorem which says that if $p(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ is conical in $X^{c\alpha}$ for some G with dim $\mathfrak{g}^{2\alpha} > 1$, then the same is true for any such G (see Theorem 8.6). Furthermore, the above metatheorems have analogues for the case dim $\mathfrak{g}^{2\alpha} = 1$, enabling us even to transfer theorems about conical vectors from any one G with dim $\mathfrak{g}^{2\alpha} = 1$ to any G with *either* dim $\mathfrak{g}^{2\alpha} = 1$ *or* dim $\mathfrak{g}^{2\alpha} > 1$ (see Theorems 8.4 and 8.6).

The conical vectors still had to be computed for some special G with dim $\mathfrak{g}^{2\alpha} \geq 1$. The only cases which we were able to do directly, aided by a crucial observation of L. Corwin, were the cases $G = SU(n, 1)$—essentially all the G's such that dim $\mathfrak{g}^{2\alpha} = 1$. In these cases, $(\mathfrak{N}^-)^\mathfrak{m}$ is the polynomial algebra in $q_{-\alpha}$ and $r_{-2\alpha}$, where $r_{-2\alpha}$ is a nonzero element of the one-dimensional space $\mathfrak{g}^{-2\alpha}$. We reformulated the condition that $p(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ be conical in X^ν (where p is a complex polynomial in two variables) in terms of a complicated system of linear equations whose unknowns were essentially the coefficients of p. These equations implied uniqueness of the conical vectors, but it was not clear that the equations had a consistent solution (and hence it was not clear that the conical vectors in Theorem 1.1 existed) until Corwin noticed that a solution vector could be constructed from the coefficients of a certain polynomial which factored into certain linear factors. This meant that if p were this polynomial, then $p(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ would be conical. This was enough to prove Theorem 1.1 for these G's. To place the case dim $\mathfrak{g}^{2\alpha} = 1$ in perspective, we further note the following: In this case, $r_{-2\alpha}^2 = q_{-2\alpha}$ in $\mathfrak{N}^-$, and therefore the factors $q_{-\alpha}^2 + j^2 q_{-2\alpha}$ in Theorem 1.1 themselves factor into linear factors: $(q_{-\alpha} + (-1)^{1/2} j r_{-2\alpha})(q_{-\alpha} - (-1)^{1/2} j r_{-2\alpha})$. It was this which made it feasible to carry out the necessary computations (see the Remark following Lemma 9.1).

Actually, in writing up the special case in §9, we dealt only with $G = SU(2, 1)$, and following a suggestion of N. Wallach, we used the theory of Verma modules to prove the uniqueness of the conical vectors. (For $G = SU(2, 1)$, the $\mathfrak{g}$-module induced from $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ is actually a Verma module, not just a quotient.) Thus the original approach, using the complicated system of linear equations, is not carried out in this paper.

The above results are stated for G of real rank 1, but they imply a result for arbitrary real rank, included in Theorems 10.1 and 10.2.

There is another direction in which Theorems 1.1 and 1.2 are extended in this paper—to arbitrary fields of characteristic zero. In fact, throughout this paper, we work with semisimple symmetric Lie algebras with splitting Cartan subspaces, over fields of characteristic zero (see [2] and [7(b)] for background on these). This accounts for most of the length of §§2–4, in which we wanted to give a self-contained elementary treatment of the Kostant-Mostow double transitivity theorem and its consequences for algebras of polynomial invariants, valid over general fields of characteristic zero, without using any theory of Lie or algebraic groups. Instead of group orbits, we use "infinitesimal transitivity and double transitivity" conditions. We essentially give Wallach's modified version of Kostant's proof of the double transitivity theorem. See §§3 and 4 for a more detailed discussion of this theorem and its consequences.

Incidentally, it is not surprising that theorems about real semisimple Lie algebras, Cartan decompositions and Iwasawa decompositions should also hold for more general semisimple symmetric Lie algebras, since joint work with G. McCollum has shown that assertions about such structures whose truth is preserved under field extension and restriction are true for any one field of characteristic zero if and only if they are true for any other; see [8(e)]. This gives a generalization of H. Weyl's "unitary trick", which enables one to transfer theorems from compact semisimple Lie algebras to semisimple Lie algebras over arbitrary fields of characteristic zero.

After the work for this paper was completed, we found a simpler proof of the uniqueness of the conical vectors, avoiding the use of the double transitivity theorem; see [8(d)]. (But the existence and explicit form of the conical vectors still require the fundamental commutation relation and transfer principles.) This proof uses an observation of Kostant on the limitations imposed on conical vectors by the action of the center of $\mathfrak{G}$. The proof also uses an a priori argument that the first assertion of Theorem 1.2 holds—that $\dim \operatorname{Hom}_{\mathfrak{g}}(X^{\mu}, X^{\nu}) \leq 1$. In fact, we have generalized this last inequality to all parabolic subalgebras (see [8(c)]) by extending the method that Verma originally used (see [2, Théorème 7.6.6]) to prove the corresponding fact about Verma modules.

We remarked above that a $\mathfrak{g}$-module X induced from a finite-dimensional irreducible $(\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n})$-module is a quotient of a certain Verma module V, but that one cannot very well use V to determine the highest weight vectors in X. On the other hand, since Theorems 1.1 and 1.2 *are* true, we can use *them* as a tool in investigating the composition series of the Verma module V. Interesting things happen: First, recall that in [1(a)], Bernšteĭn, Gel'fand and

Gel'fand found an example of a Verma module for $\mathfrak{sl}(4, \mathbf{C})$ having two strange properties: It contains a proper submodule not generated by Verma submodules, and its composition series contains a certain irreducible subquotient with multiplicity two. But it now turns out that if one regards $\mathfrak{sl}(4, \mathbf{C})$ as the complexification of $\mathfrak{su}(3, 1)$, then one can explain all of this pathology by means of the existence of a certain conical vector in X which does not come from a highest weight vector in V. In effect, Bernšteĭn, Gel'fand and Gel'fand were actually dealing with the case $l = 1$, $\zeta_l = q_{-\alpha}$ in Theorem 1.1. Moreover, using Theorem 1.1, we can generate whole families of examples of the same two "strange" phenomena for many Lie algebras. Thus a "bad" phenomenon for Verma modules becomes "good" when one interprets the situation using a larger parabolic subalgebra than a Borel subalgebra. This further emphasizes the importance of studying modules induced from general parabolic subalgebras.

Along the same lines, we comment that the results of [1] and [10] do not, in general, give explicit expressions for the highest weight vectors in a Verma module, or equivalently, explicit formulas for the embedding of one Verma module into another; they usually give only the *existence* of the vectors or the embeddings. But we can use the polynomials ζ_l in Theorem 1.1 to give explicit expressions for certain of these highest weight vectors or embeddings which have not yet been described explicitly.

We would like to thank G. D. Mostow for informing us about his approach to the double transitivity theorem.

Notations. We shall write $\mathbf{Z}_+$ for the set of nonnegative integers and $\mathbf{Q}$ for the field of rational numbers. Throughout this paper, k is a field of characteristic zero. The dual of a vector space V over k is denoted V^*. The symmetric algebra of V is written $S(V)$, and for all $r \in \mathbf{Z}_+$, the rth symmetric power is denoted $S^r(V)$, so that

$$S(V) = \coprod_{r \in \mathbf{Z}_+} S^r(V).$$

$S(V^*)$ is naturally isomorphic to the algebra of polynomial functions on V (i.e., the algebra of sums of products of linear functions on V), and we shall often identify these two algebras. Let $\mathfrak{g}$ be a Lie algebra over k, and let V be a $\mathfrak{g}$-module. Then $\mathfrak{g}$ may be canonically embedded in the universal enveloping algebra $\mathfrak{G}$ of $\mathfrak{g}$, and V may be regarded naturally as a $\mathfrak{G}$-module. The action of $\mathfrak{G}$ on V will be denoted $x \cdot v$ ($x \in \mathfrak{G}$, $v \in V$). If $\mathfrak{s}$ and T are subsets of $\mathfrak{g}$ and V, respectively, let $T^{\mathfrak{s}}$ be the set of $\mathfrak{s}$-invariants in T, i.e., $\{t \in T \mid s \cdot t = 0 \text{ for all } s \in \mathfrak{s}\}$. Regard $\mathfrak{G}$ and $S(\mathfrak{g})$ as $\mathfrak{g}$-modules by the natural extensions by derivations of the adjoint action of $\mathfrak{g}$ on itself. Then for

$x \in \mathfrak{g}$ and $y \in \mathfrak{G}$, $x \cdot y = [x, y]$, where we use $[\cdot, \cdot]$ to denote the commutator in associative algebras, as well as the bracket in Lie algebras. In particular, if $\mathfrak{s} \subset \mathfrak{g}$ and $T \subset \mathfrak{G}$, then $T^{\mathfrak{s}}$ is the ordinary centralizer of $\mathfrak{s}$ in T. Note that for all $x \in \mathfrak{g}$, $y \in \mathfrak{G}$ and $v \in V$, we have $x \cdot (y \cdot v) = [x, y] \cdot v + y \cdot (x \cdot v)$. Regard V^* as the $\mathfrak{g}$-module contragredient to the $\mathfrak{g}$-module V.

2. **The setting.** Here we shall summarize the necessary preliminaries and fix notation to be used throughout most of this paper.

Let $(\mathfrak{g}, \theta)$ be a semisimple symmetric Lie algebra over k, i.e., $\mathfrak{g}$ is a semisimple Lie algebra over k and θ is an automorphism of $\mathfrak{g}$ such that $\theta^2 = 1$. (See [2] and [7(b)] for background information on semisimple symmetric Lie algebras.) Denote by $\mathfrak{k}$ and $\mathfrak{p}$ the $+1$ and -1 eigenspaces for θ, so that $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ is the symmetric decomposition of $(\mathfrak{g}, \theta)$, orthogonal with respect to the Killing form of $\mathfrak{g}$. Assume that there is a splitting Cartan subspace $\mathfrak{a}$ of $\mathfrak{p}$. That is, $\mathfrak{a}$ is a maximal abelian subspace of $\mathfrak{p}$ whose adjoint action on $\mathfrak{g}$ can be simultaneously diagonalized.

Let $\mathfrak{m}$ be the centralizer of $\mathfrak{a}$ in $\mathfrak{k}$, and for all k-linear functionals $\phi: \mathfrak{a} \to k$, define

$$\mathfrak{g}^{\phi} = \{x \in \mathfrak{g} \mid [a, x] = \phi(a)x \text{ for all } a \in \mathfrak{a}\}.$$

Then $\mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a}$. Let

$$\Sigma = \{\phi \in \mathfrak{a}^* \mid \phi \neq 0 \text{ and } \mathfrak{g}^{\phi} \neq 0\},$$

the set of *restricted roots* of $\mathfrak{g}$ with respect to $\mathfrak{a}$. Then

$$\mathfrak{g} = \mathfrak{g}^0 \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^{\phi} = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^{\phi}.$$

Moreover, $[\mathfrak{g}^{\phi}, \mathfrak{g}^{\psi}] \subset \mathfrak{g}^{\phi \oplus \psi}$ and $\theta \mathfrak{g}^{\phi} = \mathfrak{g}^{-\phi}$ for all $\phi, \psi \in \mathfrak{a}^*$.

Let B be the Killing form of $\mathfrak{g}$. Then B is nonsingular on $\mathfrak{a}$ (see [7(b)]), so that B induces naturally a nonsingular symmetric k-bilinear form $(\cdot, \cdot)$ on $\mathfrak{a}^*$, as well as a natural isometry between $\mathfrak{a}$ and $\mathfrak{a}^*$. Let $\mathfrak{a}^*_{\mathbf{Q}}$ denote the rational span of Σ in $\mathfrak{a}^*$. Then $\mathfrak{a}^*$ is naturally isomorphic to $\mathfrak{a}^*_{\mathbf{Q}} \otimes_{\mathbf{Q}} k$, and the form $(\cdot, \cdot)$ is rational-valued and positive definite on the rational space $\mathfrak{a}^*_{\mathbf{Q}}$ (see [7(b)]). In particular, $(\phi, \phi) \neq 0$ for all $\phi \in \Sigma$.

For all $\phi \in \Sigma$, let s_{ϕ} denote the orthogonal reflection of $\mathfrak{a}^*$ through the hyperplane perpendicular to ϕ, and let W be the group of isometries of $\mathfrak{a}^*$ generated by the s_{ϕ} $(\phi \in \Sigma)$. W is called the *restricted Weyl group* of $\mathfrak{g}$ with respect to $\mathfrak{a}$. Σ spans $\mathfrak{a}^*$ and forms a (not necessarily reduced) system of roots in $\mathfrak{a}^*$ with Weyl group W (see [7(b), §2]).

Let Σ_+ be a positive system in Σ, and define

$$\mathfrak{n} = \coprod_{\phi \in \Sigma_+} \mathfrak{g}^{\phi} \quad \text{and} \quad \mathfrak{n}^- = \coprod_{\phi \in \Sigma_+} \mathfrak{g}^{-\phi}.$$

Then $\mathfrak{n}$ and $\mathfrak{n}^-$ are nilpotent subalgebras of $\mathfrak{g}$, and we have the decomposition $\mathfrak{g} = \mathfrak{n}^- \oplus \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$.

Define the bilinear form B_θ on $\mathfrak{g}$ by the condition $B_\theta(x, y) = -B(x, \theta y)$ $(x, y \in \mathfrak{g})$. Then B_θ is a nonsingular symmetric form, and the decomposition $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{a} \oplus \coprod_{\phi \in \Sigma} \mathfrak{g}^{\phi}$ is a B_θ-orthogonal decomposition (see [7(b), Lemma 3.2]). Hence B_θ is nonsingular on each $\mathfrak{g}^{\phi}$ $(\phi \in \Sigma)$ on $\mathfrak{m}$ and on $\mathfrak{a}$. Moreover, B_θ is clearly a $\mathfrak{k}$-invariant and θ-invariant form on $\mathfrak{g}$.

For all $\phi \in \Sigma$, let $x_\phi \in \mathfrak{a}$ denote the image of ϕ under the canonical isometry from $\mathfrak{a}^*$ to $\mathfrak{a}$, so that $B(x_\phi, a) = \phi(a)$ for all $a \in \mathfrak{a}$, and $B(x_\phi, x_\psi) = (\phi, \psi)$ for all $\phi, \psi \in \Sigma$. Then for all $e \in \mathfrak{g}^{\phi}$, $[e, \theta e] \in \mathfrak{a}$, and in fact

$$[e, \theta e] = B(e, \theta e) x_\phi = -B_\theta(e, e) x_\phi$$

[7(b), Lemma 3.3]. Since $(\phi, \phi) \neq 0$, we can define $h_\phi = 2x_\phi/(\phi, \phi) \in \mathfrak{a}$. Then $\phi(h_\phi) = 2$.

Suppose now that k is algebraically closed, so that every element in k has a square root. Since B_θ is a symmetric nonsingular form on $\mathfrak{g}^{\phi}$, $\mathfrak{g}^{\phi}$ contains a nonisotropic vector e_0 with respect to the form B_θ (i.e., $B_\theta(e_0, e_0) \neq 0$). Set

$$e_\phi = (2/(\phi, \phi) B_\theta(e_0, e_0))^{1/2} e_0$$

and $f_\phi = -\theta e_\phi$. Then $B_\theta(e_\phi, e_\phi) = 2/(\phi, \phi)$, and so $[h_\phi, e_\phi] = 2e_\phi$, $[h_\phi, f_\phi] = -2f_\phi$ and $[e_\phi, f_\phi] = h_\phi$. Hence $\{h_\phi, e_\phi, f_\phi\}$ spans a three-dimensional simple subalgebra $\mathfrak{u}_\phi$ of $\mathfrak{g}$.

Now drop the algebraic closure assumption on k. Let $\mathfrak{G}$ be the universal enveloping algebra of $\mathfrak{g}$, and let $\mathfrak{M}, \mathfrak{A}, \mathfrak{N}$ and $\mathfrak{N}^-$ denote the universal enveloping algebras of $\mathfrak{m}$, $\mathfrak{a}$, $\mathfrak{n}$ and $\mathfrak{n}^-$, respectively, regarded as canonically embedded in $\mathfrak{G}$. Then the multiplication map in $\mathfrak{G}$ induces a linear isomorphism

$$\mathfrak{G} \simeq \mathfrak{N}^- \otimes \mathfrak{M} \otimes \mathfrak{A} \otimes \mathfrak{N}.$$

Let $\nu \in \mathfrak{a}^*$. Then the linear form on the subalgebra $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ of $\mathfrak{g}$ which is ν on $\mathfrak{a}$ and zero on $\mathfrak{m} \oplus \mathfrak{n}$ vanishes on the commutator subalgebra of $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$, and thus corresponds to a one-dimensional representation π of $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ and hence of its universal enveloping algebra $\mathfrak{M}\mathfrak{A}\mathfrak{N}$. Let V^ν be the $\mathfrak{g}$-module induced by the $(\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n})$-module defined by π (see [2, §5.1]). That is,

$$V^\nu = \mathfrak{G} \otimes_{\mathfrak{M}\mathfrak{A}\mathfrak{N}} k,$$

where $\mathfrak{G}$ is regarded as a right $\mathfrak{M}\mathfrak{A}\mathfrak{N}$-module by right multiplication, and k is regarded as the $\mathfrak{M}\mathfrak{A}\mathfrak{N}$-module defined by π. The vector $v_0 = 1 \otimes 1 \in V^\nu$ generates V^ν as a $\mathfrak{G}$-module, and is called the *canonical generator* of V^ν. It is clear that the map $\omega: \mathfrak{N}^- \to V^\nu$ given by $x \mapsto x \cdot v_0$ is a linear isomorphism.

Let V be a $\mathfrak{g}$-module, $v \in V$ a nonzero vector and $\lambda \in \mathfrak{a}^*$. Then v is called a *restricted weight vector* and λ a *restricted weight* for V if $x \cdot v = \lambda(x)v$ for all $x \in \mathfrak{a}$. For all $\lambda \in \mathfrak{a}^*$, the subspace of V consisting of 0 and the restricted weight vectors for λ is called the *restricted weight space* for λ; it is nonzero if and only if λ is a restricted weight for V.

The following definitions are central to this paper: Let V be a $\mathfrak{g}$-module, and let $v \in V$ be nonzero. Then v is a *conical vector* for V if $v \in V^{\mathfrak{m}\oplus\mathfrak{n}}$, i.e., if $(\mathfrak{m} \oplus \mathfrak{n}) \cdot v = 0$. The subspace $V^{\mathfrak{m}\oplus\mathfrak{n}}$ consisting of 0 and the conical vectors is called the *conical space* of V.

Now let $\nu \in \mathfrak{a}^*$ and let v_0 be the canonical generator of the induced module V^ν. Then v_0 is clearly a conical restricted weight vector in V^ν with restricted weight ν. It is also clear that the conical space of V^ν is $\mathfrak{a}$-invariant and hence is the direct sum of its intersections with the restricted weight spaces of V^ν.

The standard universal property of the induced module V^ν (see [2, §5.1]) say that if U is a $\mathfrak{g}$-module and $u \in U$ is a conical restricted weight vector with restricted weight ν, then there is a unique $\mathfrak{g}$-module homomorphism $f: V^\nu \to U$ such that $f(v_0) = u$. If u generates U, then f is surjective. If $U = V^\mu$ for some $\mu \in \mathfrak{a}^*$, then f is injective; this follows from the fact that $\mathfrak{N}^-$ has no zero divisors. Let $Z \subset V^\mu$ be the intersection of the conical space and the restricted weight space for ν. Then we have a natural linear isomorphism

$$\mathrm{Hom}_{\mathfrak{g}}(V^\nu, V^\mu) \to Z, \qquad f \mapsto f(v_0).$$

Let ν and v_0 be as above. Since $v_0 \in (V^\nu)^{\mathfrak{m}}$, the linear isomorphism $\omega: \mathfrak{N}^- \to V^\nu$ (see above) is also an $\mathfrak{m}$-module isomorphism, where $\mathfrak{N}^-$ is regarded as an $\mathfrak{m}$-submodule of $\mathfrak{G}$ under the adjoint action. In particular, $(V^\nu)^{\mathfrak{m}} = (\mathfrak{N}^-)^{\mathfrak{m}} \cdot v_0$, and in fact ω restricts to a linear isomorphism

$$\omega: (\mathfrak{N}^-)^{\mathfrak{m}} \to (V^\nu)^{\mathfrak{m}}, \qquad x \mapsto x \cdot v_0.$$

Define $\rho \in \mathfrak{a}^*$ by the formula

$$\rho(a) = \tfrac{1}{2}\mathrm{tr}(\mathrm{ad}\, a \mid \mathfrak{n})$$

for all $a \in \mathfrak{a}$, i.e.,

$$\rho = \frac{1}{2} \sum_{\phi \in \Sigma_+} (\dim \mathfrak{g}^{\phi})\phi.$$

For all $\nu \in \mathfrak{a}^*$, define the $\mathfrak{g}$-module X^{ν} to be the induced module $V^{\nu-\rho}$. As above, let π be the one-dimensional representation of $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ defined by ν. Then X^{ν} can be interpreted as the twisted induced module induced by the one-dimensional $(\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n})$-module corresponding to π, in the sense of [2, §5.2]. That is, for all $m \in \mathfrak{m}$, $a \in \mathfrak{a}$ and $n \in \mathfrak{n}$, the trace of the action of $m + a + n$ on $\mathfrak{g}/(\mathfrak{m}\oplus\mathfrak{a}\oplus\mathfrak{n})$ is $-\mathrm{tr}(\mathrm{ad}\, a|\mathfrak{n}) = -2\rho(a)$. But we shall not need this fact.

The *canonical linear isomorphism* $\lambda: S(\mathfrak{g}) \to \mathcal{G}$ is defined by the formula

$$\lambda(g_1 \cdots g_n) = \frac{1}{n!} \sum_{\sigma} g_{\sigma(1)} \cdots g_{\sigma(n)}$$

for all $n \in \mathbf{Z}_+$ and $g_i \in \mathfrak{g}$; here the product on the left is taken in $S(\mathfrak{g})$, the products on the right are taken in $\mathcal{G}$, and σ ranges through the group of permutations of $\{1, \ldots, n\}$ (see [2, §2.4]). For all $g \in \mathfrak{g}$ and $n \in \mathbf{Z}_+$, $\lambda(g^n) = g^n$. Also, λ is a $\mathfrak{g}$-module isomorphism (see [2, §2.4.10]).

Let $\overline{k}$ be a field extension of k, $\overline{\mathfrak{g}} = \mathfrak{g} \otimes_k \overline{k}$, $\overline{\mathfrak{k}} = \mathfrak{k} \otimes_k \overline{k}$, etc., and let $\overline{\theta}$ be the $\overline{k}$-linear extension of θ to $\overline{\mathfrak{g}}$. Then $(\overline{\mathfrak{g}}, \overline{\theta})$ is a semisimple symmetric Lie algebra over $\overline{k}$ with symmetric decomposition $\overline{\mathfrak{g}} = \overline{\mathfrak{k}} \oplus \overline{\mathfrak{p}}$, $\overline{\mathfrak{a}}$ is a splitting Cartan subspace of $\overline{\mathfrak{p}}$, etc. We shall often use the technique of extension to a "sufficiently large" field $\overline{k}$, which can always be taken to be an algebraic closure of k. For example, the construction of the subalgebra $\mathfrak{u}_{\phi}$ above might have to be carried out over an extension field $\overline{k}$ of k, but results about $(\overline{\mathfrak{g}}, \overline{\theta})$ proved using $\mathfrak{u}_{\phi}$ can often be transferred to $(\mathfrak{g}, \theta)$.

3. **General results on polynomial invariants.** Let U be a finite-dimensional real Euclidean space and $SO(U)$ the rotation group of U. There is a natural $SO(U)$-invariant quadratic element t of the second symmetric power $S^2(U^*)$ given by the sum of the squares of the members of the dual basis of any orthonormal basis of U (t is the "square of the radius"). Let I be the algebra of $SO(U)$-invariant polynomial functions on U, or equivalently, the algebra of $SO(U)$-invariants in the symmetric algebra $S(U^*)$. A standard result of classical invariant theory states that I is exactly the set of polynomials in t if $\dim U > 1$. (If $\dim U = 1$, then $SO(U)$ acts trivially on U, and so $I = S(U^*)$.)

Clearly, I is exactly the set of polynomial functions on U constant on the $SO(U)$-orbits in U, i.e., the spheres centered at the origin if $\dim U > 1$, and the points if $\dim U = 1$. If M is any Lie group which acts as isometries

on U in such a way that M acts transitively on the $SO(U)$-orbits in U (i.e., the M-orbits in U are the same as the $SO(U)$-orbits), then the set of M-invariant polynomial functions on U must coincide with the set I of $SO(U)$-invariants.

Now suppose that M also acts as isometries on a second finite-dimensional Euclidean space V so that M acts transitively on the $SO(V)$-orbits in V. Then the set of M-invariant polynomial functions on V is the set $J \subset S(V^*)$ of $SO(V)$-invariants, and J is a polynomial algebra as above.

Now M and $SO(U) \times SO(V)$ both act naturally on $U \oplus V$. Let L be the set of M-invariants in $S((U \oplus V)^*) = S(U^*) \otimes S(V^*)$. It is easy to see that the set of $SO(U) \times SO(V)$-invariants in $S((U \oplus V)^*)$ is exactly $I \otimes J$, and that $I \otimes J \subset L$. It is important to know that $I \otimes J = L$ in certain situations. In this case, for example, L will be a polynomial algebra on two generators. In order to insure this, it is natural to assume that the M-orbits in $U \oplus V$ are the same as the $SO(U) \times SO(V)$-orbits, i.e., the products of the $SO(U)$-orbits in U with the $SO(V)$-orbits in V. This assumption is equivalent to the "double-transitivity" hypothesis—that if A is an $SO(U)$-orbit in U and B is an $SO(V)$-orbit in V, then the isotropy group of M at any point of A acts transitively on B. If $\dim U > 1$ and $\dim V > 1$, this is equivalent to saying that M acts transitively on the product of the unit sphere in U with the unit sphere in V. Under the double transitivity hypothesis, $L = I \otimes J$.

The present section is devoted to algebraic analogues of these facts, valid over the field k of characteristic zero, assumed for convenience to be algebraically closed throughout this section. Here we are concerned with a Lie algebra $\mathfrak{m}_0$ (over k) which acts on modules U and V with nonsingular symmetric $\mathfrak{m}_0$-invariant bilinear forms. Replacing the orbit hypotheses for M by corresponding "infinitesimal transitivity and double-transitivity" assumptions, we show that the $\mathfrak{m}_0$-invariant polynomial functions on U, V and $U \oplus V$ are exact analogues of the spaces of M-invariants above. We also transfer these results to the symmetric algebras $S(U)$, $S(V)$ and $S(U \oplus V) = S(U) \otimes S(V)$; the invariants here are essentially the same as for the spaces of polynomial functions. We do not need any theory of algebraic groups. The setup in this section is entirely independent of §2; the results here will be applied to the setting of §2 in the next section.

Let $\mathfrak{m}_0$ be a Lie algebra over k, U a nonzero finite-dimensional $\mathfrak{m}_0$-module, and B_0 a nonsingular symmetric $\mathfrak{m}_0$-invariant bilinear form on U. The homogeneous quadratic polynomial function $x \mapsto B_0(x, x)$ on U defines a canonical nonzero element $t_0 \in S^2(U^*)^{\mathfrak{m}_0}$ under the natural identification between the algebra of polynomial functions on U and $S(U^*)$. B_0 also induces

a canonical $\mathfrak{m}_0$-module isomorphism $\xi_0 : U^* \to U$ which extends to an $\mathfrak{m}_0$-module and algebra isomorphism $\xi_0 : S(U^*) \to S(U)$. Let $p_0 = \xi_0(t_0)$, so that $p_0 \in S^2(U)^{\mathfrak{m}_0}$.

For every element $e \in U$, denote by $e^{\perp}$ the B_0-orthogonal complement of e in U. Recall that e is called *isotropic* (resp., *nonisotropic*) with respect to B_0 if $B_0(e, e) = 0$ (resp., $B_0(e, e) \neq 0$). Note that e is B_0-nonisotropic if and only if $U = ke \oplus e^{\perp}$.

Lemma 3.1. *For all* $e \in U$, $\mathfrak{m}_0 \cdot e \subset e^{\perp}$.

Proof. Let $x \in \mathfrak{m}_0$. Then $B_0(x \cdot e, e) = -B_0(e, x \cdot e) = -B_0(x \cdot e, e)$ since B_0 is $\mathfrak{m}_0$-invariant and symmetric, and so $B_0(x \cdot e, e) = 0$. Q.E.D.

We now make the key assumption that for every B_0-nonisotropic vector $e \in U$, we have $\mathfrak{m}_0 \cdot e = e^{\perp}$. This can be thought of as an "infinitesimal transitivity" hypothesis. Our goal now is to compute $S(U)^{\mathfrak{m}_0}$, and in fact to prove:

Theorem 3.2. *If* $\dim U = 1$, *then* $S(U)^{\mathfrak{m}_0} = S(U)$. *If* $\dim U \geq 2$, *then* $S(U)^{\mathfrak{m}_0}$ *is the polynomial algebra generated by* p_0. *In particular,* $S(U)^{\mathfrak{m}_0}$ *is a polynomial algebra on one generator.*

The proof will be carried out in a series of lemmas. First we settle the easy one-dimensional case:

Lemma 3.3. *Suppose* $\dim U = 1$. *Then* $\mathfrak{m}_0$ *acts trivially on* U. *In particular,* $S(U)^{\mathfrak{m}_0} = S(U)$.

Proof. Any nonzero element e of U is B_0-nonisotropic, and so $e^{\perp} = 0$. Thus $\mathfrak{m}_0 \cdot e = 0$ (Lemma 3.1). Q.E.D.

It is also convenient to handle the two-dimensional case separately:

Lemma 3.4. *Suppose* $\dim U = 2$. *Then* $S(U)^{\mathfrak{m}_0}$ *is the polynomial algebra generated by* p_0.

Proof. Since k is algebraically closed, we may choose a B_0-orthonormal basis $\{e_1, e_2\}$ of U. Then $p_0 = e_1^2 + e_2^2 \in S^2(U)$. By hypothesis, there exists $x \in \mathfrak{m}_0$ such that $x \cdot e_1 = e_2$. Since B_0 is $\mathfrak{m}_0$-invariant, we have $B_0(e_1, x \cdot e_2) = -B_0(x \cdot e_1, e_2) = -B_0(e_2, e_2) = -1$. But $x \cdot e_2$ is a multiple of e_1, and so $x \cdot e_2 = -e_1$.

Again since k is algebraically closed, k contains a square root i of -1. Let $v_1 = e_1 + ie_2$ and $v_2 = e_1 - ie_2$, so that $\{v_1, v_2\}$ is a basis of U. Then $x \cdot v_1 = e_2 - ie_1 = -iv_1$ and $x \cdot v_2 = e_2 + ie_1 = iv_2$.

Let $f \in S(U)^{\mathfrak{m}_0}$. Then f is a polynomial of the form

$$f = \sum_{\alpha,\,\beta\in Z_+} c_{\alpha\beta} v_1^\alpha v_2^\beta$$

in v_1 and v_2 $(c_{\alpha\beta} \in k)$. Since $x \cdot f = 0$, we have

$$\sum_{\alpha,\,\beta\in Z_+} ic_{\alpha\beta}(\beta - \alpha) v_1^\alpha v_2^\beta = 0,$$

so that $c_{\alpha\beta} = 0$ unless $\alpha = \beta$. Thus $f = \Sigma_{\alpha\in Z_+} c_{\alpha\alpha}(v_1 v_2)^\alpha$. But $v_1 v_2 = e_1^2 + e_2^2 = p_0$, and so f is a polynomial in p_0. Conversely, it is clear that any polynomial in p_0 is in $S(U)^{\mathfrak{m}_0}$. The lemma now follows from the fact that the subalgebra of $S(U)$ generated by p_0 is isomorphic to the polynomial algebra generated by p_0. Q.E.D.

In order to compute $S(U)^{\mathfrak{m}_0}$ in general, we shall use the following result:

Lemma 3.5. *Let $e \in U$ be B_0-nonisotropic, and let $r \in \mathbf{Z}_+$. Then e^r generates $S^r(U)$ as an $\mathfrak{m}_0$-module. In particular,*

$$S^r(U) = ke^r + \mathfrak{m}_0 \cdot S^r(U).$$

Proof. The second statement clearly follows from the first, and so it is sufficient to prove by induction on $j = 0, \dots, r$ that the smallest $\mathfrak{m}_0$-invariant subspace T of $S^r(U)$ containing e^r also contains $e^{r-j}S^j(e^\perp)$. This is clearly true for $j = 0$, so assume it is true for $0, \dots, j$ $(j < r)$. Let $x \in \mathfrak{m}_0$ and $s \in S^j(e^\perp)$. Then

$$x \cdot e^{r-j}s = (r-j)e^{r-(j+1)}(x \cdot e)s + e^{r-j}(x \cdot s).$$

The left-hand side and the second term on the right are in T by the induction hypothesis, and so $e^{r-(j+1)}(x \cdot e)s \in T$ since $r - j > 0$. The lemma now follows from the assumption that $\mathfrak{m}_0 \cdot e = e^\perp$. Q.E.D.

The point is the following:

Lemma 3.6. *Let $e \in U$ be B_0-nonisotropic, $r \in \mathbf{Z}_+$ and $f \in S^r(U^*)^{\mathfrak{m}_0}$. Regard f as a polynomial function on U. Then f is determined by its value at e. Equivalently, if $f(e) = 0$, then $f = 0$.*

Proof. There is a natural pairing $\{\cdot,\cdot\}$ between $S^r(U^*)$ and $S^r(U)$ given as follows:

$$\{f_1 \cdots f_r, u_1 \cdots u_r\} = \sum_\sigma \prod_{i=1}^r \langle f_i, u_{\sigma(i)}\rangle,$$

where $u_1, \dots, u_r \in U$, $f_1, \dots, f_r \in U^*$, $\langle\cdot,\cdot\rangle$ is the natural pairing between U^* and U and σ ranges through the group of permutations of $\{1, \dots, r\}$. Then $\{f, u^r\} = r!f(u)$ for all $f \in S^r(U^*)$ and $u \in U$, where f is regarded as a

polynomial function on U on the right-hand side. It follows that $\{\cdot,\cdot\}$ is nonsingular. Also, the natural actions of $\mathfrak{m}_0$ on $S^r(U^*)$ and $S^r(U)$ are contragredient with respect to $\{\cdot,\cdot\}$ (see for example the proof of [7(b), Lemma 3.6]).

Now let f and e be as in the statement of the lemma. If $f(e) = 0$, then $\{f, e^r\} = 0$. Since f is $\mathfrak{m}_0$-invariant, $\{f, x \cdot s\} = -\{x \cdot f, s\} = 0$ for all $x \in \mathfrak{m}_0$ and $s \in S^r(U)$. Thus $\{f, S^r(U)\} = 0$ by Lemma 3.5, and so $f = 0$ by the nonsingularity of $\{\cdot,\cdot\}$. Q.E.D.

Theorem 3.2 now follows by applying the canonical isomorphism $\xi_0 : S(U^*) \to S(U)$ to the following result:

Lemma 3.7. *Let* $\dim U \geq 3$. *Then* $S(U^*)^{\mathfrak{m}_0}$ *is the polynomial algebra generated by* t_0. *Equivalently, if* $r \in \mathbf{Z}_+$ *is odd, then* $S^r(U^*)^{\mathfrak{m}_0} = 0$, *and if* $r = 2m$, $m \in \mathbf{Z}_+$, *then* $S^r(U^*)^{\mathfrak{m}_0}$ *is spanned by* t_0^m.

Proof. Since $S(U^*)^{\mathfrak{m}_0}$ is the direct sum of its homogeneous components, it is sufficient to compute $S^r(U^*)^{\mathfrak{m}_0}$ for $r \in \mathbf{Z}_+$. Let $V \subset U$ be the algebraic set defined by the equation $t_0(v) = 0$ $(v \in U)$. Then V is exactly the set of B_0-isotropic vectors in U. Let $f \in S^r(U^*)^{\mathfrak{m}_0}$. If f has a zero outside V, then $f = 0$ by Lemma 3.6. Hence we may assume that all the zeros of f lie in V. But then by the Hilbert Nullstellensatz, f divides some power of t_0. Choose a B_0-orthonormal basis of U, and let $X_1, \ldots, X_n \in U^*$ be the corresponding dual basis. Then $S(U^*)$ can be identified with the polynomial algebra $k[X_1, \ldots, X_n]$, and $t_0 = X_1^2 + \cdots + X_n^2$. Since $\dim U \geq 3$, t_0 is an irreducible polynomial. The fact that f divides a power of t_0 thus implies that f is itself a power of t_0 up to a scalar multiple. Q.E.D.

Theorem 3.2 is now proved.

Remark. The last assertion of Lemma 3.7 (the case $r = 2m$) can also be proved more directly (even when $\dim U \leq 2$) as follows: Let $f \in S^r(U^*)^{\mathfrak{m}_0}$, let $e \in U$ be a B_0-nonisotropic vector, and set $c = (t_0^m)(e) = t_0(e)^m \in k$. Since $t_0(e) = B_0(e, e) \neq 0$, we have $c \neq 0$. But $f(e)t_0^m$ and cf are two elements of $S^r(U^*)^{\mathfrak{m}_0}$ which take the same value $cf(e)$ at e. Hence $f = c^{-1}f(e)t_0^m$, by Lemma 3.6, proving the assertion.

The following consequence is interesting, but it will not be needed:

Corollary 3.8 (to Theorem 3.2). *Every* $\mathfrak{m}_0$*-invariant symmetric bilinear form on* U *is a scalar multiple of* B_0.

Proof. From Theorem 3.2, $S^2(U)^{\mathfrak{m}_0} = kp_0$, and so $S^2(U^*)^{\mathfrak{m}_0} = kt_0$. The corollary now follows by polarization. Q.E.D.

Remark. Corollary 3.8 has a direct proof which does not use either Lemma 3.4 or Lemma 3.6: Let C be an $\mathfrak{m}_0$-invariant symmetric bilinear form on U. Then the unique linear operator $A: U \to U$ defined by $C(u, v) = B_0(Au, v)$ for all $u, v \in U$ is an $\mathfrak{m}_0$-module map which is symmetric with respect to B_0. Let $e \in U$ be a B_0-nonisotropic vector, and let $e' \in e^{\perp}$. By hypothesis, there exists $x \in \mathfrak{m}_0$ such that $x \cdot e = e'$. Then

$$B_0(Ae, e') = B_0(Ae, x \cdot e) = -B_0(x \cdot Ae, e) = -B_0(A(x \cdot e), e)$$

$$= -B_0(Ae', e) = -B_0(e, Ae') = -B_0(Ae, e'),$$

and so $B_0(Ae, e') = 0$. Thus every B_0-nonisotropic vector of U is an eigenvector for A. Since every two B_0-orthogonal B_0-nonisotropic vectors have a B_0-nonisotropic linear combination not proportional to either of them, we see that they must have the same eigenvalue for A. Applying this to a B_0-orthogonal basis of U consisting of B_0-nonisotropic vectors shows that A is a scalar, and this completes the proof.

Another general result is required for the next section. Let V be a nonzero finite-dimensional $\mathfrak{m}_0$-module with a nonsingular symmetric $\mathfrak{m}_0$-invariant bilinear form B_1. Let $p_1 \in S^2(V)^{\mathfrak{m}_0}$ be the corresponding canonical invariant. The symmetric algebra of the direct sum $\mathfrak{m}_0$-module $U \oplus V$ is naturally isomorphic to $S(U) \otimes S(V)$, and $\mathfrak{m}_0$ acts on $S(U \oplus V)$ according to the tensor product of its actions on $S(U)$ and $S(V)$. In particular, $S(U)^{\mathfrak{m}_0} \otimes S(V)^{\mathfrak{m}_0} \subset S(U \oplus V)^{\mathfrak{m}_0}$. The next theorem gives an important case in which this inclusion becomes an equality.

Theorem 3.9. *In the context of Theorem 3.2, suppose in addition that for every B_0-nonisotropic vector $e_0 \in U$ and every B_1-nonisotropic vector $e_1 \in V$, we have $\mathfrak{m}_0' \cdot e_1 = e_1^{\perp}$ in V, where $\mathfrak{m}_0'$ is the centralizer of e_0 in $\mathfrak{m}_0$. Then*

$$S(U \oplus V)^{\mathfrak{m}_0} = S(U)^{\mathfrak{m}_0} \otimes S(V)^{\mathfrak{m}_0},$$

$S(U)^{\mathfrak{m}_0}$ is given by Theorem 3.2, and $S(V)^{\mathfrak{m}_0}$ is either $S(V)$ or the polynomial algebra generated by p_1, depending on whether $\dim V = 1$ or $\dim V \geq 2$. In particular, $S(U \oplus V)^{\mathfrak{m}_0}$ is a polynomial algebra on two generators.

Proof. Let $e_0 \in U$ be B_0-nonisotropic, and let $\mathfrak{m}_0'$ be the centralizer of e_0 in $\mathfrak{m}_0$. For every B_1-nonisotropic vector $e_1 \in V$, we have $e_1^{\perp} = \mathfrak{m}_0' \cdot e_1 \subset \mathfrak{m}_0 \cdot e_1 \subset e_1^{\perp}$ by Lemma 3.1, so that $\mathfrak{m}_0 \cdot e_1 = e_1^{\perp}$. Thus Theorem 3.2 applies to $\mathfrak{m}_0$, V, B_1 and p_1, and so to prove the theorem all we must show is that $S(U \oplus V)^{\mathfrak{m}_0} \subset S(U)^{\mathfrak{m}_0} \otimes S(V)^{\mathfrak{m}_0}$.

We shall now apply a technique used in [7(b), §5]. It is clear that

$S(U \oplus V)^{\mathfrak{m}_0}$ is the direct sum of its homogeneous components of the form $(S^r(U) \otimes S(V))^{\mathfrak{m}_0}$, where $r \in \mathbf{Z}_+$, and so it is sufficient to show that $(S^r(U) \otimes S(V))^{\mathfrak{m}_0} \subset S^r(U)^{\mathfrak{m}_0} \otimes S(V)^{\mathfrak{m}_0}$.

Recall the nonsingular $\mathfrak{m}_0$-invariant pairing $\{\cdot,\cdot\}$ between $S^r(U^*)$ and $S^r(U)$ (see the proof of Lemma 3.6). Also recall the canonical $\mathfrak{m}_0$-module and algebra isomorphism $\xi_0 : S(U^*) \to S(U)$. Then ξ_0 restricts to an $\mathfrak{m}_0$-module isomorphism $\xi_0 : S^r(U^*) \to S^r(U)$. Define a bilinear map

$$\omega : S^r(U) \otimes S(V) \times S^r(U) \to S(V)$$

by the condition $s \otimes w, t \mapsto \{\xi_0^{-1}(s), t\}w$ for all $s, t \in S^r(U)$ and $w \in S(V)$. Then for all $x \in \mathfrak{m}_0$, $y \in S^r(U) \otimes S(V)$ and $t \in S^r(U)$, we have

$$\omega(x \cdot y, t) + \omega(y, x \cdot t) = x \cdot \omega(y, t).$$

Moreover, let X be any subspace of $S(V)$. We claim that for all $y \in S^r(U) \otimes S(V)$, $\omega(y, S^r(U)) \subset X$ implies $y \in S^r(U) \otimes X$. In fact, choose a basis $\{w_i\}$ for a complement of X in $S^r(V)$ and write $y = \sum_i s_i \otimes w_i + z$ $(s_i \in S^r(U)$, $z \in S^r(U) \otimes X)$. Then for all $t \in S^r(U)$, we have

$$\sum_i \omega(s_i \otimes w_i, t) + \omega(z, t) \in X,$$

and so $\sum_i \{\xi_0^{-1}(s_i), t\}w_i \in X$. Hence $\{\xi_0^{-1}(s_i), S^r(U)\} = 0$ for all i, so that each $s_i = 0$, proving the claim.

Let $y \in (S^r(U) \otimes S(V))^{\mathfrak{m}_0}$, and let e_0 and $\mathfrak{m}_0'$ be as in the statement of the theorem. Then for all $x \in \mathfrak{m}_0'$,

$$x \cdot \omega(y, e_0^r) = \omega(x \cdot y, e_0^r) + \omega(y, r(x \cdot e_0)e_0^{r-1}) = 0$$

since $x \cdot y = 0$ and $x \cdot e_0 = 0$. Hence $\omega(y, e_0^r) \in S(V)^{\mathfrak{m}_0'}$. But by hypothesis, $\mathfrak{m}_0' \cdot e_1 = e_1^{\perp}$ in V, for every B_1-nonisotropic vector $e_1 \in V$. Thus Theorem 3.2 applies to $\mathfrak{m}_0'$, V, B_1 and p_1, as well as to $\mathfrak{m}_0$, V, B_1 and p_1. In particular, $S(V)^{\mathfrak{m}_0'} = S(V)^{\mathfrak{m}_0}$, and so $\omega(y, e_0^r) \in S(V)^{\mathfrak{m}_0}$. But the set Z of B_0-nonisotropic vectors in U is Zariski dense since it is the set on which the polynomial function $t_0 \in S^2(U^*)$ does not vanish. Hence the powers e_0^r $(e_0 \in Z)$ span $S^r(U)$ (see for example [7(b), Lemma 3.5(ii)]). It follows that $\omega(y, S^r(U)) \subset S(V)^{\mathfrak{m}_0}$. But now the above claim applied to $X = S(V)^{\mathfrak{m}_0}$ implies that $y \in S^r(U) \otimes S(V)^{\mathfrak{m}_0}$.

The rest is easy: Let $\{a_i\}$ be a basis of $S(V)^{\mathfrak{m}_0}$, and write $y = \sum_i b_i \otimes a_i$ $(b_i \in S^r(U))$. Since $\mathfrak{m}_0 \cdot y = 0$, we must have $\sum_i x \cdot b_i \otimes a_i = 0$ for all $x \in \mathfrak{m}_0$, so that $\mathfrak{m}_0 \cdot b_i = 0$ for each i. Hence $y \in S^r(U)^{\mathfrak{m}_0} \otimes S(V)^{\mathfrak{m}_0}$, and the theorem is proved. Q.E.D.

4. The Kostant-Mostow double transitivity theorem. In this section, we return to the setting of §2. For every $\phi \in \Sigma$, $\mathfrak{m}$ acts naturally on the subalgebra $\mathfrak{n}_\phi = \mathfrak{g}^\phi \oplus \mathfrak{g}^{2\phi}$ of $\mathfrak{g}$. (Here $\mathfrak{g}^{2\phi}$ might be zero.) Our main goal at this point is to determine the algebra $S(\mathfrak{n}_\phi)^\mathfrak{m}$ of $\mathfrak{m}$-invariants in the symmetric algebra $S(\mathfrak{n}_\phi)$. It will turn out to be a polynomial algebra on one or two generators (Theorem 4.6). The method will be to verify the hypotheses of §3 and then to apply the results of §3.

Suppose that $\dim \mathfrak{a} = 1$, ϕ is the unique simple restricted root, $\dim \mathfrak{g}^{2\phi} > 1$, $k = \mathbf{R}$, θ is a Cartan involution of $\mathfrak{g}$ in the sense that the Killing form of $\mathfrak{g}$ is negative definite on $\mathfrak{k}$ and positive definite on $\mathfrak{p}$, G is a connected Lie group corresponding to $\mathfrak{g}$, K is the connected Lie subgroup of G corresponding to $\mathfrak{k}$, and M is the centralizer of $\mathfrak{a}$ in K. Then $S(\mathfrak{n}_\phi)^\mathfrak{m}$ is the space $S(\mathfrak{n}_\phi)^M$ of M-invariants in $S(\mathfrak{n}_\phi)$, and determining $S(\mathfrak{n}_\phi)^M$ amounts to proving a double transitivity theorem for the action of M on $\mathfrak{g}^\phi \oplus \mathfrak{g}^{2\phi}$. Specifically, let S_1 be the unit sphere in $\mathfrak{g}^\phi$, and S_2 the unit sphere in $\mathfrak{g}^{2\phi}$, with respect to the bilinear form B_θ, which is positive definite on $\mathfrak{g}$. The issue is to prove that M acts transitively on $S_1 \times S_2$. This theorem was proved by B. Kostant [6, §2.1] (in a somewhat different formulation) and independently by G. D. Mostow (oral communication; related ideas are discussed in [8, §19]). Kostant's proof, as modified slightly by N. Wallach [11, Theorem 8.11.3], is purely algebraic. In order to show that this proof applies in our general setting, and for our later reference, we shall give an exposition of Kostant's proof below. (Mostow's proof is based on explicit case-by-case checking; only the case of the exceptional group F_4 is difficult.) We have been discussing the rather subtle situation in which $\dim \mathfrak{g}^{2\phi} > 1$; if $\dim \mathfrak{g}^{2\phi} \leq 1$, the appriate results are very easy.

Return now to the general setting of §2.

Fix $\phi \in \Sigma$. We shall describe a canonical element $p_\phi \in S^2(\mathfrak{g}^\phi)^\mathfrak{m}$. The symmetric bilinear form B_θ is nonsingular on $\mathfrak{g}^\phi$ (see §2). Since B_θ is $\mathfrak{k}$-invariant and hence $\mathfrak{m}$-invariant, and since $\mathfrak{g}^\phi$ is $\mathfrak{m}$-stable, the restriction of B_θ to $\mathfrak{g}^\phi$ is $\mathfrak{m}$-invariant. As in §3, we get a nonzero homogeneous quadratic polynomial function $x \mapsto B_\theta(x, x)$ on $\mathfrak{g}^\phi$, and this defines a nonzero element $t_\phi \in S^2((\mathfrak{g}^\phi)^*)^\mathfrak{m}$. B_θ induces a canonical $\mathfrak{m}$-module isomorphism $\xi_\phi : (\mathfrak{g}^\phi)^* \to \mathfrak{g}^\phi$ which extends to an $\mathfrak{m}$-module and algebra isomorphism $\xi_\phi : S((\mathfrak{g}^\phi)^*) \to S(\mathfrak{g}^\phi)$. Let $p_\phi = \xi_\phi(t_\phi)$, so that $p_\phi \in S^2(\mathfrak{g}^\phi)^\mathfrak{m}$.

Now we shall verify that the key assumption of the beginning of §3 holds in the present context, with $\mathfrak{m}_0 = \mathfrak{m}$ acting on $U = \mathfrak{g}^\phi$ by the adjoint action, and $B_0 = B_\theta | \mathfrak{g}^\phi \times \mathfrak{g}^\phi$. The word "nonisotropic" and the symbol $e^\perp$ have the meanings of §3.

Lemma 4.1 (cf. [6, Theorem 2.1.7]). *Let $e_0 \in \mathfrak{g}^{\phi}$ be a B_θ-nonisotropic vector. Then $[\mathfrak{m}, e_0] = e_0^{\perp}$. In particular, $\mathfrak{g}^{\phi} = ke_0 \oplus [\mathfrak{m}, e_0]$.*

Proof. It is clearly sufficient to assume that k is algebraically closed. As in §2, we may choose a multiple e_ϕ of e_0 such that $B_\theta(e_\phi, e_\phi) = 2/(\phi, \phi)$. Setting $h_\phi = 2x_\phi/(\phi, \phi) \in \mathfrak{a}$ and $f_\phi = -\theta e_\phi$, we have the bracket relations $[h_\phi, e_\phi] = 2e_\phi$, $[h_\phi, f_\phi] = -2f_\phi$ and $[e_\phi, f_\phi] = h_\phi$ (see §2), so that $\{h_\phi, e_\phi, f_\phi\}$ spans a three-dimensional simple Lie subalgebra $\mathfrak{u}_\phi$ of $\mathfrak{g}$. Let $\mathfrak{g}_\phi$ be the $\mathfrak{u}_\phi$-submodule $\amalg_{j=-2}^{2} \mathfrak{g}^{j\phi}$ of $\mathfrak{g}$. Since the eigenspaces of ad h_ϕ in $\mathfrak{g}_\phi$ with eigenvalues 0 and 2 are $\mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a}$ and $\mathfrak{g}^{\phi}$, respectively, the representation theory of a three-dimensional simple Lie algebra implies that $[e_\phi, \mathfrak{m} \oplus \mathfrak{a}] = \mathfrak{g}^{\phi}$. But $[e_\phi, \mathfrak{m}] \subset e_\phi^{\perp}$ by Lemma 3.1, and since $[e_\phi, \mathfrak{a}] = ke_\phi$, we must have $[e_\phi, \mathfrak{m}] = e_\phi^{\perp}$. The lemma is now clear. Q. E. D.

Before applying Theorem 3.2, we shall derive two more results:

Lemma 4.2. *We have $[\mathfrak{g}^{\phi}, \mathfrak{g}^{\phi}] = \mathfrak{g}^{2\phi}$.*

Proof. We may assume that k is algebraically closed. As in §2 (or the last proof), we have the three-dimensional simple Lie subalgebra $\mathfrak{u}_\phi$ of $\mathfrak{g}$ spanned by h_ϕ, e_ϕ and f_ϕ. Let $\mathfrak{g}_\phi$ be the $\mathfrak{u}_\phi$-submodule $\amalg_{j=-2}^{2} \mathfrak{g}^{j\phi}$ of $\mathfrak{g}$. The eigenspaces of h_ϕ in $\mathfrak{g}_\phi$ with eigenvalues 2 and 4 are $\mathfrak{g}^{\phi}$ and $\mathfrak{g}^{2\phi}$, respectively, and so the representation theory of $\mathfrak{u}_\phi$ implies that $[e_\phi, \mathfrak{g}^{\phi}] = \mathfrak{g}^{2\phi}$. Q. E. D.

The following consequence will be useful later:

Corollary 4.3. *Let X be a $\mathfrak{g}$-module and $x \in X$ an $\mathfrak{m}$-invariant vector annihilated by some B_θ-nonisotropic vector $e_0 \in \mathfrak{g}^{\phi}$. Then $(\mathfrak{g}^{\phi} \oplus \mathfrak{g}^{2\phi}) \cdot x = 0$. In particular, if* dim $\mathfrak{a} = 1$ *and ϕ is the unique simple restricted root, then x is a conical vector in X.*

Proof. For all $y \in \mathfrak{m}$, $[y, e_0] \cdot x = y \cdot (e_0 \cdot x) - e_0 \cdot (y \cdot x) = 0$, and so $\mathfrak{g}^{\phi} \cdot x = 0$ by Lemma 4.1. Lemma 4.2 now implies that $\mathfrak{g}^{2\phi} \cdot x = 0$. The last assertion is clear. Q. E. D.

Theorem 3.2, Lemma 4.1 and the field extension technique imply:

Theorem 4.4. *If* dim $\mathfrak{g}^{\phi} = 1$, *then $S(\mathfrak{g}^{\phi})^{\mathfrak{m}} = S(\mathfrak{g}^{\phi})$. If* dim $\mathfrak{g}^{\phi} \geq 2$, *then $S(\mathfrak{g}^{\phi})^{\mathfrak{m}}$ is the polynomial algebra generated by p_ϕ. In particular, $S(\mathfrak{g}^{\phi})^{\mathfrak{m}}$ is a polynomial algebra on one generator.*

Corollary 4.5. *Every $\mathfrak{m}$-invariant symmetric bilinear form on $\mathfrak{g}^{\phi}$ is a scalar multiple of B_θ.*

The corollary follows from either Theorem 4.4 or Corollary 3.8; see the

remark following Corollary 3.8 for a simple proof. We shall not have to use Corollary 4.5.

Our next goal is to verify the hypothesis of Theorem 3.9 for $U = \mathfrak{g}^{\phi}$ and $V = \mathfrak{g}^{2\phi}$ in case $2\phi \in \Sigma$ (see Lemma 4.7). This amounts to proving the Kostant-Mostow double transitivity theorem. For reasons mentioned above, we shall essentially repeat Kostant's proof [6, §2.1], with a couple of modifications (the proofs of Lemmas 4.18 and 4.20) taken from Wallach's exposition [11, Theorem 8.11.3]. The result is:

Theorem 4.6. *Suppose $\phi \in \Sigma$, and let $\mathfrak{n}_{\phi}$ be the subalgebra $\mathfrak{g}^{\phi} \oplus \mathfrak{g}^{2\phi}$ of $\mathfrak{g}$. Then $S(\mathfrak{n}_{\phi})^{\mathfrak{m}} = S(\mathfrak{g}^{\phi})^{\mathfrak{m}} \otimes S(\mathfrak{g}^{2\phi})^{\mathfrak{m}}$, and this is a polynomial algebra. Moreover, let $p_{\phi} \in S^2(\mathfrak{g}^{\phi})^{\mathfrak{m}}$ be the canonical quadratic $\mathfrak{m}$-invariant defined by B_{θ}, and if $2\phi \in \Sigma$, let $p_{2\phi} \in S^2(\mathfrak{g}^{2\phi})^{\mathfrak{m}}$ be the same for 2ϕ. Then there are four possibilities:*

Case 1. dim $\mathfrak{g}^{\phi} = 1$ *and* $\mathfrak{g}^{2\phi} = 0$. *Let $x \in \mathfrak{g}^{\phi}$, $x \neq 0$. Then $S(\mathfrak{n}_{\phi})^{\mathfrak{m}} = S(\mathfrak{g}^{\phi}) = k[x]$, and $k[x]$ is the polynomial algebra generated by x.*

Case 2. dim $\mathfrak{g}^{\phi} > 1$ *and* $\mathfrak{g}^{2\phi} = 0$. *Then $S(\mathfrak{n}_{\phi})^{\mathfrak{m}} = k[p_{\phi}]$, and $k[p_{\phi}]$ is the polynomial algebra generated by p_{ϕ}.*

Case 3. dim $\mathfrak{g}^{\phi} >$ dim $\mathfrak{g}^{2\phi} = 1$. *Then $S(\mathfrak{g}^{\phi})^{\mathfrak{m}} = k[p_{\phi}]$ and $S(\mathfrak{g}^{2\phi})^{\mathfrak{m}} = S(\mathfrak{g}^{2\phi}) = k[y]$, where $y \in \mathfrak{g}^{2\phi}$, $y \neq 0$. Both algebras are polynomial algebras in the indicated generators, so that $S(\mathfrak{n}_{\phi})^{\mathfrak{m}}$ is the polynomial algebra $k[p_{\phi}, y]$ in the two generators p_{ϕ} and y.*

Case 4. dim $\mathfrak{g}^{\phi} >$ dim $\mathfrak{g}^{2\phi} > 1$. *Then $S(\mathfrak{g}^{\phi})^{\mathfrak{m}}$ and $S(\mathfrak{g}^{2\phi})^{\mathfrak{m}}$ are the polynomial algebras $k[p_{\phi}]$ and $k[p_{2\phi}]$, respectively, so that $S(\mathfrak{n}_{\phi})^{\mathfrak{m}}$ is the polynomial algebra $k[p_{\phi}, p_{2\phi}]$.*

Proof. We may, and do, assume that k is algebraically closed. The fact that dim $\mathfrak{g}^{\phi} >$ dim $\mathfrak{g}^{2\phi}$ will be proved in Lemma 4.8. Also, Cases 1 and 2 are covered in Theorem 4.4. The rest of Theorem 4.6 follows immediately from Theorem 3.9, Lemma 4.1 and:

Lemma 4.7. *Suppose $\phi, 2\phi \in \Sigma$. Let $e_0 \in \mathfrak{g}^{\phi}$ and $e_1 \in \mathfrak{g}^{2\phi}$ be B_{θ}-nonisotropic, and let $\mathfrak{m}_0$ be the centralizer of e_0 in $\mathfrak{m}$. Then $[\mathfrak{m}_0, e_1] = e_1^{\perp}$ in $\mathfrak{g}^{2\phi}$.*

This result will follow from the next series of lemmas. Note that only Case 4 of Theorem 4.6 remains to be proved, since Lemma 4.7 is trivial if dim $\mathfrak{g}^{2\phi} = 1$. But it will not be necessary in the following proof to impose any restriction on dim $\mathfrak{g}^{2\phi}$, and in fact the proof holds even if $\mathfrak{g}^{2\phi} = 0$.

We shall use the notation of the proof of Lemma 4.1, so that e_{ϕ} is a certain multiple of e_0, and $\{h_{\phi}, e_{\phi}, f_{\phi}\}$ spans a three-dimensional simple

subalgebra $\mathfrak{u}_\phi$ of $\mathfrak{g}$. Also as in the proof of Lemma 4.1, let $\mathfrak{g}_\phi$ be the $\mathfrak{u}_\phi$-submodule $\coprod_{j=-2}^{2} \mathfrak{g}^{j\phi}$ of $\mathfrak{g}$. The natural representation of $\mathfrak{u}_\phi$ on $\mathfrak{g}_\phi$ decomposes $\mathfrak{g}_\phi$ into a direct sum of irreducible $\mathfrak{u}_\phi$-submodules. Since the eigenvalues of ad h_ϕ on $\mathfrak{g}_\phi$ are among 0, ±2 and ±4 (with corresponding eigenspaces $\mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a}$, $\mathfrak{g}^{\pm\phi}$ and $\mathfrak{g}^{\pm 2\phi}$), the dimensions of the irreducible components can only be 1, 3 and 5. A five-dimensional irreducible module occurs if and only if $\mathfrak{g}^{2\phi} \neq 0$, and a three-dimensional irreducible module always occurs—$\mathfrak{u}_\phi$ itself. Let $\mathfrak{g}_i \subset \mathfrak{g}_\phi$ be the sum of all the $(2i+1)$-dimensional irreducible $\mathfrak{u}_\phi$-submodules of $\mathfrak{g}_\phi$ $(i = 0, 1, 2)$, so that $\mathfrak{g}_\phi = \mathfrak{g}_0 \oplus \mathfrak{g}_1 \oplus \mathfrak{g}_2$. Also, let $\mathfrak{g}_i^j = \mathfrak{g}_i \cap \mathfrak{g}^{j\phi}$ $(0 \le i \le 2, -2 \le j \le 2)$; then $\mathfrak{g}_i = \coprod_{j=-i}^{i} \mathfrak{g}_i^j$ for each $i = 0, 1, 2$. Also, $\mathfrak{g}^{\pm 2\phi} = \mathfrak{g}_2^{\pm 2}$, $\mathfrak{g}^{\pm\phi} = \mathfrak{g}_1^{\pm 1} \oplus \mathfrak{g}_2^{\pm 1}$ and $\mathfrak{g}^0 = \mathfrak{g}_0^0 \oplus \mathfrak{g}_1^0 \oplus \mathfrak{g}_2^0$.

Lemma 4.8. *We have* dim $\mathfrak{g}^\phi >$ dim $\mathfrak{g}^{2\phi}$.

Proof. This is clear since $\mathfrak{g}^\phi = \mathfrak{g}_1^1 \oplus \mathfrak{g}_2^1$, $\mathfrak{g}^{2\phi} = \mathfrak{g}_2^2$, dim $\mathfrak{g}_2^1 =$ dim $\mathfrak{g}_2^2$ and dim $\mathfrak{g}_1^1 \ge 1$ (since $e_\phi \in \mathfrak{g}_1^1$). Q.E.D.

Lemma 4.9. *The decomposition* $\mathfrak{g}_\phi = \mathfrak{g}_0 \oplus \mathfrak{g}_1 \oplus \mathfrak{g}_2$ *is both B_θ-orthogonal and B-orthogonal.*

Proof. First we shall show that $B_\theta(\mathfrak{g}_1^1, \mathfrak{g}_2^1) = 0$. Let $x \in \mathfrak{g}_1^1$, $y \in \mathfrak{g}_2^1$. Then $y = [f_\phi, z]$ for some $z \in \mathfrak{g}^{2\phi} = \mathfrak{g}_2^2$, and so

$$B_\theta(x, y) = -B(x, \theta y) = -B(x, [-e_\phi, \theta z]) = -B([e_\phi, x], \theta z) = 0$$

since $[e_\phi, x] = 0$. Hence $B_\theta(\mathfrak{g}_1^1, \mathfrak{g}_2^1) = 0$, and similar arguments show that

$$B_\theta(\mathfrak{g}_1^{-1}, \mathfrak{g}_2^{-1}) = B_\theta(\mathfrak{g}_0^0, \mathfrak{g}_1^0) = B_\theta(\mathfrak{g}_0^0, \mathfrak{g}_2^0) = 0$$

and

$$B(\mathfrak{g}_1^1, \mathfrak{g}_2^{-1}) = B(\mathfrak{g}_1^{-1}, \mathfrak{g}_2^1) = B(\mathfrak{g}_0^0, \mathfrak{g}_0^1) = B(\mathfrak{g}_0^0, \mathfrak{g}_0^2) = 0.$$

Since $B_\theta(\mathfrak{g}^{j\phi}, \mathfrak{g}^{k\phi}) = 0$ unless $j = k$, and $B(\mathfrak{g}^{j\phi}, \mathfrak{g}^{k\phi}) = 0$ unless $j = -k$, all that remains is to show that $B_\theta(\mathfrak{g}_1^0, \mathfrak{g}_2^0) = B(\mathfrak{g}_1^0, \mathfrak{g}_2^0) = 0$. Let $u \in \mathfrak{g}_1^0, v \in \mathfrak{g}_2^0$. Then $v = [f_\phi, w]$ for some $w \in \mathfrak{g}_2^1$, so that

$$B_\theta(u, v) = -B(u, \theta v) = -B(u, [-e_\phi, \theta w])$$

$$= -B([e_\phi, u], \theta w) = B_\theta([e_\phi, u], w) = 0$$

by the above, since $[e_\phi, u] \in \mathfrak{g}_1^1$ and $w \in \mathfrak{g}_2^1$. Thus $B_\theta(\mathfrak{g}_1^0, \mathfrak{g}_2^0) = 0$. Similarly, $B(\mathfrak{g}_1^0, \mathfrak{g}_2^0) = 0$. Q.E.D.

Lemma 4.10. *Let $e \in \mathfrak{g}^\phi$ and $f \in \mathfrak{g}^{-\phi}$, and suppose $B(e, f) = 0$, or equivalently, $B_\theta(f, \theta e) = 0$ or $B_\theta(e, \theta f) = 0$. Then $[e, f] \in \mathfrak{m}$.*

Proof. Since $[e, f] \in \mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a}$ and since $\mathfrak{m}$ is the B-orthogonal complement of $\mathfrak{a}$ in $\mathfrak{g}^0$, it is sufficient to show that $B([e, f], \mathfrak{a}) = 0$. But if $h \in \mathfrak{a}$, then

$$B([e, f], h) = -B(e, [h, f]) = \phi(h)B(e, f) = 0,$$

and so the lemma is proved. Q.E.D.

Lemma 4.11. *We have* $\mathfrak{g}_2^0 \subset \mathfrak{m}$.

Proof. Every element in $\mathfrak{g}_2^0$ is of the form $[e_\phi, f]$, where $f \in \mathfrak{g}_2^{-1}$. Since $e_\phi \in \mathfrak{g}_1$, Lemma 4.9 implies that $B(e_\phi, f) = 0$. But then $[e_\phi, f] \in \mathfrak{m}$ by Lemma 4.10. Q.E.D.

Lemma 4.12. *We have* $\mathfrak{g}_1^0 = kh_\phi \oplus (\mathfrak{g}_1^0 \cap \mathfrak{m})$.

Proof. It is sufficient to show that $\mathfrak{g}_1^0 \subset kh_\phi + \mathfrak{m}$. But $\mathfrak{g}_1^0 = [e_\phi, \mathfrak{g}_1^{-1}]$ and $\mathfrak{g}_1^{-1} \subset \mathfrak{g}^{-\phi} = kf_\phi \oplus f_\phi^\perp$, where $f_\phi^\perp$ is the B_θ-orthogonal complement of f_ϕ in $\mathfrak{g}^{-\phi}$. In fact, $B_\theta(f_\phi, f_\phi) = B_\theta(e_\phi, e_\phi) \neq 0$. Hence $\mathfrak{g}_1^0 \subset kh_\phi + [e_\phi, f_\phi^\perp]$, and $[e_\phi, f_\phi^\perp] \subset \mathfrak{m}$ by Lemma 4.10. Q.E.D.

Lemma 4.13. *We have* $\mathfrak{g}_0^0 = \operatorname{Ker} \phi \oplus (\mathfrak{g}_0^0 \cap \mathfrak{m})$.

Proof. Since $\mathfrak{g}_0^0 = \mathfrak{g}_0$ is the centralizer of $\mathfrak{u}_\phi$ in $\mathfrak{g}_\phi$, $\mathfrak{g}_0^0$ is stable under θ, and so $\mathfrak{g}_0^0 = (\mathfrak{g}_0^0 \cap \mathfrak{a}) \oplus (\mathfrak{g}_0^0 \cap \mathfrak{m})$. But the centralizer of $\mathfrak{u}_\phi$ in $\mathfrak{a}$ is clearly $\operatorname{Ker} \phi$, and so $\mathfrak{g}_0^0 \cap \mathfrak{a} = \operatorname{Ker} \phi$. Q.E.D.

Let $\mathfrak{m}_i = \mathfrak{g}_i \cap \mathfrak{m} = \mathfrak{g}_i^0 \cap \mathfrak{m}$ $(i = 1, 2, 3)$, and note that $\mathfrak{m}_0$ is the centralizer of e_ϕ in $\mathfrak{m}$ and hence coincides with the subalgebra $\mathfrak{m}_0$ in the statement of Lemma 4.7. The next lemma summarizes the last three:

Lemma 4.14. *We have* $\mathfrak{g}_2^0 = \mathfrak{m}_2$, $\mathfrak{g}_1^0 = kh_\phi \oplus \mathfrak{m}_1$ *and* $\mathfrak{g}_0^0 = \operatorname{Ker} \phi \oplus \mathfrak{m}_0$. *In particular,* $\mathfrak{m} = \mathfrak{m}_0 \oplus \mathfrak{m}_1 \oplus \mathfrak{m}_2$.

For all $x \in \mathfrak{g}$, define $x^* = [e_\phi, x]$. Write x^{**} instead of $(x^*)^*$. Also, define $x_* = [f_\phi, x]$, and write x_{**} for $(x_*)_*$.

Recall the following standard fact about the representation theory of the three-dimensional simple Lie algebra $\mathfrak{u}_\phi$: Let π be a finite-dimensional irreducible representation of $\mathfrak{u}_\phi$ on the space V and let $v \in V$ be a nonzero eigenvector for $\pi(h_\phi)$. Let p be the smallest nonnegative integer j such that $\pi(f_\phi)^{j+1}(v) = 0$ and q the smallest nonnegative integer j such that $\pi(e_\phi)^{j+1}(v) = 0$. Then $\pi(f_\phi)\pi(e_\phi)(v) = (p+1)qv$ and $\pi(e_\phi)\pi(f_\phi)(v) = (q+1)pv$. This implies:

Lemma 4.15. *For all* $x \in \mathfrak{m}_2$, $(x^{**})_* = 4x^*$, $(x^*)_* = 6x$, $(x_{**})^* = 4x_*$ *and* $(x_*)^* = 6x$.

Lemma 4.16. *Let* $x, y \in \mathfrak{m}_2$. *Then* $[x, y^{**}] = [x^{**}, y] = (2/3)[y^*, x^*]$.

Proof. By Lemma 4.15,

$$[x, y^{**}] = (1/6)[(x^*)_*, y^{**}] = (1/6)[(y^{**})_*, x^*] = (2/3)[y^*, x^*].$$

Hence also

$$[x^{**}, y] = -[y, x^{**}] = -(2/3)[x^*, y^*] = (2/3)[y^*, x^*]. \quad \text{Q.E.D.}$$

Lemma 4.17. *For all* $x, y \in \mathfrak{m}_2$, $[x, y]^{**} = (2/3)[x^*, y^*]$.

Proof. We have

$$[x, y]^{**} = [x^{**}, y] + 2[x^*, y^*] + [x, y^{**}] = (2/3)[x^*, y^*]$$

by Lemma 4.16. Q.E.D.

For all $x \in \mathfrak{g}_\phi$, let x_i $(i = 0, 1, 2)$ be the component of x in $\mathfrak{g}_i$ with respect to the decomposition $\mathfrak{g}_\phi = \mathfrak{g}_0 \oplus \mathfrak{g}_1 \oplus \mathfrak{g}_2$.

Lemma 4.18. *For all* $x, y \in \mathfrak{m}_2$, $[x, y]_1 = 0$.

Proof. By Lemma 4.17, $[x, y]^{**} = (2/3)[x^*, y^*]$, so that

$$([x, y]^{**})_* = (2/3)[(x^*)_*, y^*] + (2/3)[x^*, (y^*)_*]$$

$$= 4[x, y^*] + 4[x^*, y] = 4[x, y]^*,$$

using Lemma 4.15. But $[x, y]^{**} = ([x, y]_2)^{**}$, and $(([x, y]_2)^{**})_* = 4([x, y]_2)^*$ by Lemma 4.15. Hence $[x, y]^* = ([x, y]_2)^*$. But since $[x, y]^* = ([x, y]_1)^* + ([x, y]_2)^*$, we get $([x, y]_1)^* = 0$, and so $[x, y]_1 = 0$. Q.E.D.

Lemma 4.19. *For all* $x, y \in \mathfrak{m}_2$, $[[x, y]_0, y^{**}] = -2[[x, y]_2, y^{**}]$.

Proof. By Lemma 4.16, we have

$$[[x, y]_2, y^{**}] = [([x, y]_2)^{**}, y] = [[x, y]^{**}, y] = -[[x, y^{**}], y]$$

(Lemmas 4.16 and 4.17)

$$= -[[x, y], y^{**}] - [x, [y^{**}, y]] = -[[x, y], y^{**}]$$

(Lemma 4.16)

$$= -[[x, y]_0, y^{**}] - [[x, y]_2, y^{**}],$$

by Lemma 4.18. The lemma now follows. Q.E.D.

If $x \in \mathfrak{m}$, note that $x_* = -\theta x^*$ and $x_{**} = \theta x^{**}$.

Lemma 4.20. *Let* $x, y \in \mathfrak{m}_2$, *and suppose* $B_\theta(x^{**}, y^{**}) = 0$. *Then*

$[x^{**}, y_{**}] \in \mathfrak{m}$, $[x^{**}, y_{**}] = -[y^{**}, x_{**}]$, *and* $[x^{**}, y_{**}]_1 = 0$.

Proof. By Lemma 4.10 applied to $\mathfrak{g}^{2\phi}$ in place of $\mathfrak{g}^{\phi}$, $e = x^{**}$ and $f = y_{**} = \theta y^{**}$, we have $[x^{**}, y_{**}] \in \mathfrak{m}$. Thus

$$[x^{**}, y_{**}] = \theta[x^{**}, y_{**}] = [\theta x^{**}, \theta y_{**}] = [x_{**}, y^{**}] = -[y^{**}, x_{**}],$$

proving the second assertion.

To prove the last, first note that $(y_{**})^* = 4y_*$, by Lemma 4.15. Hence

$$\begin{aligned}[x^{**}, y_{**}]^* &= [x^{**}, (y_{**})^*] = 4[x^{**}, y_*] \\ &= 4[x^{**}, y]_* - 4[(x^{**})_*, y] = 4[x^{**}, y]_* - 16[x^*, y]\end{aligned}$$

(again by Lemma 4.15)

$$= -4([x, y]^{**})_* - 16[x^*, y],$$

by Lemmas 4.16 and 4.17. Thus

$$([x^{**}, y_{**}]_1)^* = -16[x^*, y]_1.$$

Hence by the second assertion, we also have

$$([x^{**}, y_{**}]_1)^* = -([y^{**}, x_{**}]_1)^* = 16[y^*, x]_1 = -16[x, y^*]_1.$$

Thus

$$([x^{**}, y_{**}]_1)^* = -8([x^*, y] + [x, y^*])_1 = -8([x, y]^*)_1 = -8([x, y]_1)^* = 0,$$

by Lemma 4.18. It is finally clear that $[x^{**}, y_{**}]_1 = 0$. Q.E.D.

Lemma 4.21. *Let* $x, y \in \mathfrak{m}_2$, *and suppose* $B_\theta(x^{**}, y^{**}) = 0$. *Then* $[x^*, y_{**}] = -6[x, y]_*$.

Proof. We have $[x_*, y^{**}] = (1/4)[x_{**}, y^{**}]^*$ by Lemma 4.15, and this is $(1/4)[x^{**}, y_{**}]^*$ by Lemma 4.20. But $[x^{**}, y_{**}] \in \mathfrak{m}$ (Lemma 4.20). Thus the last assertion of Lemma 4.20 shows that $[x_*, y^{**}] \in \mathfrak{g}_2^1$. Now

$$[x_*, y^{**}]^* = [(x_*)^*, y^{**}] = 6[x, y^{**}]$$

(by Lemma 4.15)

$$= -6[x, y]^{**}$$

by Lemmas 4.16 and 4.17. But both $[x_*, y^{**}]$ and $[x, y]^*$ are in $\mathfrak{g}_2^1$ by the above and Lemma 4.18. Hence $[x_*, y^{**}] = -6[x, y]^*$, and the lemma follows by applying θ. Q.E.D.

Lemma 4.22. *Let* $x, y \in \mathfrak{m}_2$, *and suppose* $B_\theta(x^{**}, y^{**}) = 0$ *and* $B_\theta(y^{**}, y^{**}) = 1/2(\phi, \phi)$. *Then*

$$[[x, y]_0, y^{**}] = -x^{**}/18.$$

Proof. By Lemma 4.19, $[[x, y]_0, y^{**}] = -2[[x, y]_2, y^{**}]$. But $[x, y]_2 = (1/6)([x, y]_*)^*$, by Lemmas 4.15 and 4.18, and so

$$[[x, y]_0, y^{**}] = -(1/3)[([x, y]_*)^*, y^{**}]$$

$$= -(1/3)[[x, y]_*, y^{**}]^* = (1/18)[[x^*, y_{**}], y^{**}]^*,$$

by Lemma 4.21. Also,

$$B_\theta(y^{**}, y^{**}) = 1/2(\phi, \phi) = 2/(2\phi, 2\phi),$$

and so as in §2 we must have the bracket relations for a three-dimensional simple Lie algebra, say $\mathfrak{u}_{2\phi}$, spanned by $h_{2\phi}$, y^{**} and $-\theta y^{**}$:

$$[h_{2\phi}, y^{**}] = 2y^{**}, \quad [h_{2\phi}, -\theta y^{**}] = 2\theta y^{**} \quad \text{and} \quad [y^{**}, -\theta y^{**}] = h_{2\phi}.$$

But $-\theta y^{**} = -y_{**}$ and $h_{2\phi} = \frac{1}{2} h_\phi$. Thus x^* is an eigenvector for ad $h_{2\phi}$ with eigenvalue 1, and must lie in a two-dimensional irreducible $\mathfrak{u}_{2\phi}$-submodule of $\mathfrak{g}$. Hence applying the discussion preceding Lemma 4.15 to $\mathfrak{u}_{2\phi}$, we get

$$[y^{**}, [-y_{**}, x^*]] = x^*.$$

Thus $[[x, y]_0, y^{**}] = -x^{**}/18$, and the lemma is proved. Q.E.D.

In the notation of Lemma 4.7, a multiple e' of the nonisotropic vector $e_1 \in \mathfrak{g}^{2\phi}$ may be chosen so that $B_\theta(e', e') = 1/2(\phi, \phi)$. Then e' is of the form y^{**} for some $y \in \mathfrak{m}_2$. Let $e'' \in e_1^\perp$. Then $e'' = x^{**}$ for some $x \in \mathfrak{m}_2$, and so by Lemma 4.22, $[-18[x, y]_0, e'] = e''$. Thus there exists $z \in \mathfrak{m}_0$ such that $[z, e_1] = e''$. Lemma 4.7 is finally proved, and hence so is Theorem 4.6. Q.E.D.

5. **The structure of $\mathfrak{N}_\phi^{\mathfrak{m}}$.** Continuing to work in the setting of §2, we shall transfer Theorem 4.6 to its "noncommutative analogue", i.e., to the structure theorem for $\mathfrak{N}_\phi^{\mathfrak{m}}$ (see below).

Retain the notation of §4. In particular, $\phi \in \Sigma$ is fixed. Recall the canonical linear isomorphism $\lambda: S(\mathfrak{g}) \to \mathfrak{G}$. Let $\mathfrak{N}_\phi$ be the universal enveloping algebra of the Lie subalgebra $\mathfrak{n}_\phi = \mathfrak{g}^\phi \oplus \mathfrak{g}^{2\phi}$ of $\mathfrak{g}$ defined in Theorem 4.6, so that $\lambda: S(\mathfrak{n}_\phi) \to \mathfrak{N}_\phi$ is an $\mathfrak{m}$-module isomorphism which restricts to a linear isomorphism from $S(\mathfrak{n}_\phi)^{\mathfrak{m}}$ to $\mathfrak{N}_\phi^{\mathfrak{m}}$. We shall now use Theorem 4.6 to give an explicit description of the algebra $\mathfrak{N}_\phi^{\mathfrak{m}}$. Recall the canonical quadratic $\mathfrak{m}$-invariants $p_\phi \in S^2(\mathfrak{g}^\phi)^{\mathfrak{m}}$ and (if $2\phi \in \Sigma$) $p_{2\phi} \in S^2(\mathfrak{g}^{2\phi})^{\mathfrak{m}}$. Define

$$q_\phi = 2\lambda(p_\phi)/(\phi, \phi) \in \mathfrak{N}_\phi^{\mathfrak{m}},$$

and similarly, if $2\phi \in \Sigma$, define

$$q_{2\phi} = 2\lambda(p_{2\phi})/(2\phi, 2\phi) = \lambda(p_{2\phi})/2(\phi, \phi) \in \mathfrak{N}_\phi^{\mathfrak{m}}.$$

Theorem 5.1. *$\mathfrak{N}_\phi^{\mathfrak{m}}$ is commutative and in fact is a polynomial algebra. More precisely, in the four cases of Theorem 4.6, we have:*

Case 1. $\mathfrak{N}_\phi^{\mathfrak{m}} = \mathfrak{N}_\phi = k[x]$, the polynomial algebra generated by an arbitrary nonzero $x \in \mathfrak{g}^\phi$.

Case 2. $\mathfrak{N}_\phi^{\mathfrak{m}} = k[q_\phi]$, the polynomial algebra generated by q_ϕ.

Case 3. $\mathfrak{N}_\phi^{\mathfrak{m}} = k[q_\phi, y]$, where y is an arbitrary nonzero element of $\mathfrak{g}^{2\phi}$; this is the polynomial algebra in the indicated generators.

Case 4. $\mathfrak{N}_\phi^{\mathfrak{m}} = k[q_\phi, q_{2\phi}]$, the polynomial algebra generated by q_ϕ and $q_{2\phi}$.

Proof. Cases 1 and 2 follow immediately from the corresponding cases of Theorem 4.6, together with the fact that $\lambda: S(\mathfrak{n}_\phi) \to \mathfrak{N}_\phi$ is an algebra isomorphism since $\mathfrak{n}_\phi$ is abelian.

Since $\lambda(S(\mathfrak{n}_\phi)^{\mathfrak{m}}) = \mathfrak{N}_\phi^{\mathfrak{m}}$, Theorem 4.6 shows that the elements q_ϕ and y in Case 3 and q_ϕ and $q_{2\phi}$ in Case 4 lie in $\mathfrak{N}_\phi^{\mathfrak{m}}$. Also, since $\mathfrak{g}^{2\phi}$ is central in $\mathfrak{n}_\phi$, we see that q_ϕ commutes with y in Case 3 and $q_{2\phi}$ in Case 4.

Denote the usual filtration of the enveloping algebra $\mathfrak{N}_\phi$ by $\mathfrak{N}_0 \subset \mathfrak{N}_1 \subset \mathfrak{N}_2 \subset \cdots$, so that $\mathfrak{N}_0 = k \cdot 1$ and $\mathfrak{N}_1 = k \cdot 1 \oplus \mathfrak{n}_\phi$, and for each $r \in \mathbf{Z}_+$ let $\pi_r: \mathfrak{N}_r \to \mathfrak{N}_r/\mathfrak{N}_{r-1}$ be the canonical map. (Here we take $\mathfrak{N}_{-1} = 0$.) We also have the usual grading $S(\mathfrak{n}_\phi) = \coprod_{r=0}^\infty S^r(\mathfrak{n}_\phi)$ of $S(\mathfrak{n}_\phi)$. For each $r \in \mathbf{Z}_+$, let $\sigma_r: S^r(\mathfrak{n}_\phi) \to \mathfrak{N}_r/\mathfrak{N}_{r-1}$ be the canonical map, so that σ_r is a linear isomorphism by the Poincaré-Birkhoff-Witt theorem (see [2, Proposition 2.3.6]).

Now suppose that we are in Case 3. We claim that q_ϕ and y are algebraically independent. In fact, if not, then for some $r \in \mathbf{Z}_+$, there is an equation

$$\sum_{j=0}^{r} \sum_{i=0}^{[j/2]} a_{ij} q_\phi^i y^{j-2i} = 0,$$

where the $a_{ij} \in k$, and some $a_{ir} \neq 0$ $(i = 0, \ldots, [r/2])$; $[\cdot]$ denotes the "greatest integer" function. Thus $\sum_{i=0}^{[r/2]} a_{ir} q_\phi^i y^{r-2i} \in \mathfrak{N}_{r-1}$, so that

$$\pi_r\left(\sum_{i=0}^{[r/2]} a_{ir} q_\phi^i y^{r-2i}\right) = 0.$$

Consider the element

$$s = \sum_{i=0}^{[r/2]} a_{ir}\left(\frac{2}{(\phi, \phi)} p_\phi\right)^i y^{r-2i} \in S^r(\mathfrak{n}_\phi).$$

Then

$$\sigma_r(s) = \pi_r(\lambda(s)) = \pi_r\left(\sum_{i=0}^{[r/2]} a_{ir}\lambda\left(\frac{2}{(\phi, \phi)}p_\phi\right)^i \lambda(y)^{r-2i}\right)$$

$$= \pi_r\left(\sum_{i=0}^{[r/2]} a_{ir} q_\phi^i y^{r-2i}\right) = 0,$$

and so $s = 0$. But this is a contradiction, since p_ϕ and y are algebraically independent in $S(\mathfrak{n}_\phi)$, and the claim is established. A similar argument shows that in Case 4, q_ϕ and $q_{2\phi}$ are algebraically independent.

All that remains is to show that q_ϕ and y generate $\mathfrak{N}_\phi^{\mathfrak{m}}$ in Case 3 and that q_ϕ and $q_{2\phi}$ generate $\mathfrak{N}_\phi^{\mathfrak{m}}$ in Case 4. We shall carry out the argument only for Case 3; Case 4 is similar. Assume inductively that q_ϕ and y generate $\mathfrak{N}_\phi^{\mathfrak{m}} \cap \mathfrak{N}_r$, where $r \in \mathbf{Z}_+$. (This is trivially true for $r = 0$.) Now

$$\mathfrak{N}_t = \lambda\left(\coprod_{i=0}^{t} S^i(\mathfrak{n}_\phi)\right)$$

for all $t \in \mathbf{Z}_+$. Let

$$z \in \mathfrak{N}_\phi^{\mathfrak{m}} \cap \mathfrak{N}_{r+1} = \lambda\left(S(\mathfrak{n}_\phi)^{\mathfrak{m}} \cap \coprod_{i=0}^{r+1} S^i(\mathfrak{n}_\phi)\right).$$

Then z is of the form

$$z = \lambda\left(\sum_{j=0}^{r+1}\sum_{i=0}^{[j/2]} a_{ij}\left(\frac{2}{(\phi, \phi)}p_\phi\right)^i y^{j-2i}\right)$$

$(a_{ij} \in k)$ by Theorem 4.6. But

$$z - \sum_{j=0}^{r+1}\sum_{i=0}^{[j/2]} a_{ij} q_\phi^i y^{j-2i} \in \mathfrak{N}_\phi^{\mathfrak{m}} \cap \mathfrak{N}_r,$$

and so the induction hypothesis implies that z can be expressed as a polynomial in q_ϕ and y. This completes the proof of Theorem 5.1. Q.E.D.

6. **The case $2\alpha \notin \Sigma$.** In this section, we compute certain commutators in the universal enveloping algebra $\mathfrak{G}$ of $\mathfrak{g}$, and then we use these to determine certain conical vectors in the twisted induced $\mathfrak{g}$-modules X^ν, where $\nu \in \mathfrak{a}^*$ (see §2). Specifically, we prove our main results (Theorems 10.1 and 10.2) in the special case in which twice the relevant restricted root is not a restricted root (see Theorems 6.17 and 6.18). But the first part of the section, through Lemma 6.4, is valid in general, and this will be important in §8.

Maintain the hypotheses and notation of the last section. For conven-

ience assume for awhile (through Corollary 6.10) that k is algebraically closed.

Continue to fix $\phi \in \Sigma$, and choose h_ϕ, e_ϕ, f_ϕ, and $\mathfrak{u}_\phi$ as in §2. Applying the constructions of the beginnings of §§4 and 5 to $-\phi$ in place of ϕ, we have canonical elements $p_{-\phi} \in S^2(\mathfrak{g}^{-\phi})^{\mathfrak{m}}$ and $q_{-\phi} = 2\lambda(p_{-\phi})/(\phi, \phi) \in \mathfrak{N}^{\mathfrak{m}}_{-\phi}$. Our goal now is to compute the commutator $[e_\phi, q_{-\phi}]$ in $\mathcal{G}$.

Since $B_\theta(e_\phi, e_\phi) = 2/(\phi, \phi)$, it is clear that $B_\theta(f_\phi, f_\phi) = 2(\phi, \phi)$ also. Using the notation of the proof of Lemma 4.7, we recall from Lemma 4.9 that the decomposition $\mathfrak{g}_\phi = \mathfrak{g}_0 \oplus \mathfrak{g}_1 \oplus \mathfrak{g}_2$ is B_θ-orthogonal, and hence so is the decomposition $\mathfrak{g}^{-\phi} = \mathfrak{g}_1^{-1} \oplus \mathfrak{g}_2^{-1}$. Set $f_1 = f_\phi$. Since B_θ is nonsingular on $\mathfrak{g}^{-\phi}$ and k is algebraically closed, we may complete f_1 to a B_θ-orthogonal basis $\{f_1, \ldots, f_n\}$ of $\mathfrak{g}^{-\phi}$ such that $B_\theta(f_i, f_i) = 2/(\phi, \phi)$ for all $i = 1, \ldots, n$. But since $f_1 \in \mathfrak{g}_1^{-1}$, we may also assume that $f_1, \ldots, f_r \in \mathfrak{g}_1^{-1}$ and that $f_{r+1}, \ldots, f_n \in \mathfrak{g}_2^{-1}$. Here $n = \dim \mathfrak{g}^\phi = \dim \mathfrak{g}^{-\phi}$ and $r = \dim \mathfrak{g}_1^{-1}$. Note that $\dim \mathfrak{g}^{2\phi} = \dim \mathfrak{g}_2^{-1} = n - r$, and hence that $\mathfrak{g}^{2\phi} \neq 0$ if and only if $r < n$.

The canonical element $p_{-\phi} \in S^2(\mathfrak{g}^{-\phi})$ is equal to the sum of the squares of the elements of any B_θ-orthonormal basis of $\mathfrak{g}^{-\phi}$, and so

$$p_{-\phi} = \frac{(\phi, \phi)}{2} \sum_{i=1}^{n} f_i^2.$$

Thus

$$q_{-\phi} = \frac{2}{(\phi, \phi)} \lambda(p_{-\phi}) = \sum_{i=1}^{n} f_i^2 \in \mathfrak{N}_{-\phi}.$$

To compute $[e_\phi, q_{-\phi}]$, we first note that

$$\begin{aligned}
[e_\phi, q_{-\phi}] &= \sum_{i=1}^{n} [e_\phi, f_i^2] = \sum_{i=1}^{n} ([e_\phi, f_i]f_i + f_i[e_\phi, f_i]) \\
&= \sum_{i=1}^{n} ([[e_\phi, f_i], f_i] + 2f_i[e_\phi, f_i]) \\
&= \sum_{i=1}^{n} ([f_i, [f_i, e_\phi]] + 2f_i[e_\phi, f_i]).
\end{aligned}$$

Lemma 6.1. $[f_1, [f_1, e_\phi]] = -2f_1$.

Proof. This follows immediately from the bracket relations for h_ϕ, e_ϕ and $f_1 = f_\phi$. Q.E.D.

Lemma 6.2. *For all* $i = 2, \ldots, n$, $[e_\phi, f_i] \in \mathfrak{m}$.

Proof. Apply Lemma 4.10. Q.E.D.

Lemma 6.3. *For all* $i = 2, \ldots, r$, $[f_i, [f_i, e_\phi]] = 2f_1$, *and for all* $i = r+1, \ldots, n$, $[f_i, [f_i, e_\phi]] = 6f_1$.

Proof. Let $i = 2, \ldots, n$. Then

$$[e_\phi, [f_i, [f_i, e_\phi]]] = -[e_\phi, [f_i, [e_\phi, f_i]]] = -[f_i, [e_\phi, [e_\phi, f_i]]].$$

But $[e_\phi, f_i] \in \mathfrak{m}$ (Lemma 6.2), so that

$$\theta[e_\phi, [e_\phi, f_i]] = [\theta e_\phi, [e_\phi, f_i]] = -[f_1, [e_\phi, f_i]].$$

Now we can apply the standard representation theory of the three-dimensional simple Lie algebra $\mathfrak{u}_\phi$. If $2 \le i \le r$, then $f_i \in \mathfrak{g}_1^{-1}$, and so $[f_1, [e_\phi, f_i]] = 2f_i$, and if $r+1 \le i \le n$, then $f_i \in \mathfrak{g}_2^{-1}$, and so $[f_1, [e_\phi, f_i]] = 6f_i$. Hence $[e_\phi, [e_\phi, f_i]] = -2\theta f_i$ or $-6\theta f_i$, respectively, and so

$$[e_\phi, [f_i, [f_i, e_\phi]]] = 2[f_i, \theta f_i] \quad \text{or} \quad 6[f_i, \theta f_i],$$

respectively. But $[f_i, \theta f_i] = -B_\theta(f_i, f_i)x_{-\phi}$ (see §2), and this is just h_ϕ. Thus

$$[e_\phi, [f_i, [f_i, e_\phi]]] = 2h_\phi \quad \text{or} \quad 6h_\phi,$$

respectively. But $[f_i, [f_i, e_\phi]] \in \mathfrak{g}^{-\phi}$, and so has eigenvalue -2 for ad h_ϕ. Since $[e_\phi, [f_i, [f_i, e_\phi]]]$ is a multiple of h_ϕ, the representation theory of $\mathfrak{u}_\phi$ implies that $[f_i, [f_i, e_\phi]]$ must be a multiple of f_1. Since $[e_\phi, f_1] = h_\phi$, the multiple is determined and the lemma follows. Q.E.D.

In view of these lemmas, we have

$$\begin{aligned}[e_\phi, q_{-\phi}] &= -2f_1 + 2f_1h_\phi + 2(r-1)f_1 + 6(n-r)f_1 + 2\sum_{i=2}^{n} f_i[e_\phi, f_i] \\ &= 2\left((3n-2r-2)f_1 + f_1h_\phi + \sum_{i=2}^{n} f_i[e_\phi, f_i]\right).\end{aligned}$$

Let

$$\rho_\phi = \tfrac{1}{2}((\dim \mathfrak{g}^{\phi})\phi + (\dim \mathfrak{g}^{2\phi})(2\phi)) \in \mathfrak{a}^*. \tag{1}$$

Then $\rho_\phi = \tfrac{1}{2}(n + 2(n-r))\phi = \tfrac{1}{2}(3n-2r)\phi$, and so $\rho_\phi(h_\phi) = 3n-2r$. The conclusion is:

Lemma 6.4. *Define* ρ_ϕ *as in* (1). *Then*

$$[e_\phi, q_{-\phi}] = 2\left((\rho_\phi - \phi)(h_\phi)f_\phi + f_\phi h_\phi + \sum_{i=2}^{n} f_i[e_\phi, f_i]\right).$$

We could now use the derivation law to write down an expression for $[e_\phi, q_{-\phi}^d]$, for all $d \in \mathbf{Z}_+$. In order to simplify matters, however, we shall assume at this point that $\mathfrak{g}^{2\phi} = 0$, which implies that $\mathfrak{g}^{-\phi}$ is an abelian Lie subalgebra of $\mathfrak{g}$. The much subtler general situation is deferred to subsequent sections.

Lemma 6.5. *Suppose* $2\phi \notin \Sigma$. *For all* $d \in \mathbf{Z}_+$,

$$[e_\phi, q_{-\phi}^d] = 2dq_{-\phi}^{d-1}\Big(f_\phi(h_\phi + (\rho_\phi - d\phi)(h_\phi)) + \sum_{i=2}^{n} f_i[e_\phi, f_i]\Big).$$

Proof. From Lemmas 6.4 and 6.2 and the commutativity of $\mathfrak{g}^{-\phi}$, we get

$$\begin{aligned}[e_\phi, q_{-\phi}^d] &= \sum_{j=1}^{d} q_{-\phi}^{d-j}[e_\phi, q_{-\phi}]q_{-\phi}^{j-1} \\ &= 2dq_{-\phi}^{d-1}\Big((\rho_\phi - \phi)(h_\phi)f_\phi + \sum_{i=2}^{n} f_i[e_\phi, f_i]\Big) \\ &\quad + 2f_\phi \sum_{j=1}^{d} q_{-\phi}^{d-j}h_\phi q_{-\phi}^{j-1}.\end{aligned}$$

But $q_{-\phi}$ is clearly a restricted weight vector for the action of $\mathfrak{a}$ on $\mathfrak{G}$ with restricted weight -2ϕ, and so

$$\begin{aligned}h_\phi q_{-\phi}^{j-1} &= [h_\phi, q_{-\phi}^{j-1}] + q_{-\phi}^{j-1}h_\phi \\ &= -2(j-1)\phi(h_\phi)q_{-\phi}^{j-1} + q_{-\phi}^{j-1}h_\phi = q_{-\phi}^{j-1}(-4(j-1) + h_\phi).\end{aligned}$$

Hence

$$\sum_{i=1}^{d} q_{-\phi}^{d-j}h_\phi q_{-\phi}^{j-1} = \sum_{j=1}^{d} q_{-\phi}^{d-1}(-4(j-1) + h_\phi) = q_{-\phi}^{d-1}(dh_\phi - 2d(d-1)),$$

and so

$$\begin{aligned}[e_\phi, q_{-\phi}^d] &= 2dq_{-\phi}^{d-1}\Big((\phi d) - \phi)(h_\phi)f_\phi + \sum_{i=2}^{n} f_i[e_\phi, f_i]\Big) \\ &\quad + 2dq_{-\phi}^{d-1}f_\phi h_\phi - 4dq_{-\phi}^{d-1}f_\phi(d-1) \\ &= 2dq_{-\phi}^{d-1}\Big(f_\phi(\rho_\phi(h_\phi) - 2 - 2(d-1) + h_\phi) + \sum_{i=2}^{n} f_i[e_\phi, f_i\Big) \\ &= 2dq_{-\phi}^{d-1}\Big(f_\phi(h_\phi + (\rho_\phi - d\phi)(h_\phi)) + \sum_{i=2}^{n} f_i[e_\phi, f_i]\Big). \text{ Q.E.D.}\end{aligned}$$

The following result is now immediate:

Corollary 6.6. *Suppose $\phi \in \Sigma_+$ and $2\phi \notin \Sigma$. Let X be a $\mathfrak{g}$-module and $x \in X$ a conical restricted weight vector with restricted weight $\mu \in \mathfrak{a}^*$. Then for all $d \in \mathbf{Z}_+$,*

$$e_\phi \cdot (q_{-\phi}^d \cdot x) = 2d((\mu + \rho_\phi - d\phi)(h_\phi))/_\phi q_{-\phi}^{d-1} \cdot x.$$

If $\dim \mathfrak{g}^\phi = 1$ (in which case $\mathfrak{g}^{\pm 2\phi} = 0$ automatically), we also have the following lemma and corollary:

Lemma 6.7. *Suppose* $\dim \mathfrak{g}^\phi = 1$. *Then for all $d \in \mathbf{Z}_+$,*

$$[e_\phi, f_\phi^d] = d f_\phi^{d-1}(h_\phi + (\rho_\phi - d\phi/2)(h_\phi)).$$

Proof. Since $[e_\phi, f_\phi] = h_\phi$, we have

$$[e_\phi, f_\phi^d] = \sum_{j=1}^{d} f_\phi^{d-j} h_\phi f_\phi^{j-1}.$$

But

$$h_\phi f_\phi^{j-1} = [h_\phi, f_\phi^{j-1}] + f_\phi^{j-1} h_\phi = f_\phi^{j-1}(-2(j-1) + h_\phi),$$

so that

$$[e_\phi, f_\phi^d] = f_\phi^{d-1}(dh_\phi - d(d-1)) = d f_\phi^{d-1}(h_\phi + (\rho_\phi - d\phi/2)(h_\phi)),$$

since $\rho_\phi(h_\phi) = \frac{1}{2}\phi(h_\phi) = 1$. Q.E.D.

Corollary 6.8. *Suppose $\phi \in \Sigma_+$ and* $\dim \mathfrak{g}^\phi = 1$. *Let X be a $\mathfrak{g}$-module and $x \in X$ an $\mathfrak{n}$-invariant restricted weight vector with restricted weight $\mu \in \mathfrak{a}^*$. Then for all $d \in \mathbf{Z}_+$,*

$$e_\phi \cdot (f_\phi^d \cdot x) = d((\mu + \rho_\phi - d\phi/2)(h_\phi)) f_\phi^{d-1} \cdot x.$$

Corollaries 6.6 and 6.8 imply the following two results. These have the benefit of being true even if k is not algebraically closed, as the field extension technique shows; we also use the fact that the B_θ-nonisotropic vectors in $\mathfrak{g}^\phi$ span $\mathfrak{g}^\phi$.

Corollary 6.9. *Suppose $\phi \in \Sigma_+$ and $2\phi \notin \Sigma_+$. Let X be a $\mathfrak{g}$-module and $x \in X$ a conical restricted weight vector with restricted weight $\mu \in \mathfrak{a}^*$. Then for all $e_0 \in \mathfrak{g}^\phi$ and $d \in \mathbf{Z}_+$,*

$$e_0 \cdot (q^d_{-\phi} \cdot x) = -2d((\mu + \rho_\phi - d\phi)(h_\phi))(\theta e_0)q^{d-1}_{-\phi} \cdot x.$$

Corollary 6.10. *Suppose $\phi \in \Sigma_+$ and $\dim \mathfrak{g}^\phi = 1$. Let X be a $\mathfrak{g}$-module and $x \in X$ an $\mathfrak{n}$-invariant restricted weight vector with restricted weight $\mu \in \mathfrak{a}^*$. Then for all $e_0 \in \mathfrak{g}^\phi$ and $d \in \mathbf{Z}_+$,*

$$e_0 \cdot ((\theta e_0)^d \cdot x) = -\tfrac{1}{2} B_\theta(e_0, e_0)(\phi, \phi)d((\mu + \rho_\phi - d\phi/2)(h_\phi))(\theta e_0)^{d-1} \cdot x.$$

Assume for the rest of this section that k is an arbitrary field of characteristic zero—not necessarily algebraically closed. We are now ready to prove the following basic result:

Lemma 6.11. *Suppose $\phi \in \Sigma_+$ and $2\phi \notin \Sigma_+$. Let $\nu \in \mathfrak{a}^*$, and let x_0 be the canonical generator of the twisted induced $\mathfrak{g}$-module $X^\nu = V^{\nu-\rho}$ (see §2). Set $Y = (\mathfrak{N}_{-\phi} \cdot x_0)^{\mathfrak{m} \oplus \mathfrak{n}_\phi}$ (see §5 for the definitions of $\mathfrak{N}_{-\phi}$ and $\mathfrak{n}_\phi$), and define $h'_\phi \in \mathfrak{a}$ to be h_ϕ if $\dim \mathfrak{g}^\phi > 1$ and $2h_\phi$ if $\dim \mathfrak{g}^\phi = 1$. If $(\nu-\rho+\rho_\phi)(h'_\phi)$ is not a positive even integer, then Y is the span of x_0. Suppose $(\nu - \rho + \rho_\phi)(h'_\phi) = 2l$, l a positive integer. Then Y is two-dimensional, with basis $\{x_0, f^l \cdot x_0\}$, where $f = q_{-\phi}$ if $\dim \mathfrak{g}^\phi > 1$ and f is a nonzero element of $\mathfrak{g}^{-\phi}$ if $\dim \mathfrak{g}^\phi = 1$. In this case, $f^l \cdot x_0$ is a restricted weight vector in X^ν with restricted weight $s_\phi(\nu - \rho + \rho_\phi) - \rho_\phi$ (recall from §2 that s_ϕ is the Weyl reflection with respect to ϕ).*

Proof. Since the map $\omega : \mathfrak{N}^- \to X^\nu$ which takes $y \in \mathfrak{N}^-$ to $y \cdot x_0$ is an $\mathfrak{m}$-module isomorphism (see §2), we see that

$$(\mathfrak{N}_{-\phi} \cdot x_0)^{\mathfrak{m}} = (\omega(\mathfrak{N}_{-\phi}))^{\mathfrak{m}} = \mathfrak{N}^{\mathfrak{m}}_{-\phi} \cdot x_0.$$

But by Theorem 5.1 (Cases 1 and 2), $\mathfrak{N}^{\mathfrak{m}}_{-\phi}$ is the polynomial algebra $k[f]$, where f is as in the statement of the lemma. Hence

$$(\mathfrak{N}_{-\phi} \cdot x_0)^{\mathfrak{m}} = k[f] \cdot x_0.$$

Let $u \in k[f]$, so that $u = \sum_{d=0}^\infty a_d f^d$ ($a_d \in k$, and only finitely many $a_d \neq 0$), and let e_0 be a B_θ-nonisotropic vector in $\mathfrak{g}^\phi$; if $\dim \mathfrak{g}^\phi = 1$, take $e_0 = \theta f$. Suppose $\dim \mathfrak{g}^\phi > 1$. Then by Corollary 6.9,

$$e_0 \cdot (u \cdot x_0) = -2 \sum_{d=0}^{\infty} a_d d((\nu - \rho + \rho_\phi - d\phi)(h_\phi))(\theta e_0)q^{d-1}_{-\phi} \cdot x_0,$$

and this expression is zero if and only if $a_d d((\nu - \rho + \rho_\phi - d\phi)(h_\phi)) = 0$ for all d. But this is the case if and only if $a_d = 0$ for all $d > 0$ such that $(\nu - \rho + \rho_\phi)(h_\phi) \neq 2d$. The lemma for $\dim \mathfrak{g}^\phi > 1$ now follows from Corollary 4.3; the last assertion of the lemma is clear since $q^l_{-\phi} \cdot x_0$ has restricted

weight $\nu - \rho - 2l\phi$, and

$$s_\phi(\nu - \rho + \rho_\phi) = (\nu - \rho + \rho_\phi) - (\nu - \rho + \rho_\phi)(h_\phi)\phi = \nu - \rho + \rho_\phi - 2l\phi.$$

The case $\dim \mathfrak{g}^\phi = 1$ is similar, using Corollary 6.10. Q.E.D.

Remark. Note that Lemma 6.11 holds when ϕ is not necessarily a simple restricted root, and even when $\frac{1}{2}\phi$ is a restricted root.

The situation in Lemma 6.11 simplifies nicely when $\dim \mathfrak{a} = 1$; the next result is an immediate consequence of the lemma:

Theorem 6.12. *Suppose* $\dim \mathfrak{a} = 1$ *and* $\phi \in \Sigma_+$ *is the only positive root. Let* $\nu \in \mathfrak{a}^*$. *Then the conical space* Y *of the twisted induced* $\mathfrak{g}$*-module* X^ν *is either one- or two-dimensional. Define* $h'_\phi \in \mathfrak{a}$ *to be* h_ϕ *if* $\dim \mathfrak{g}^\phi > 1$ *and* $2h_\phi$ *if* $\dim \mathfrak{g}^\phi = 1$, *and let* x_0 *be the canonical generator of* X^ν. *If* $\nu(h'_\phi)$ *is not a positive even integer, then* Y *is the span of* x_0. *Suppose* $\nu(h'_\phi) = 2l$, l *a positive integer. Then* $\dim Y = 2$, *and* Y *has basis* $\{x_0, f^l \cdot x_0\}$, *where* $f = q_{-\phi}$ *if* $\dim \mathfrak{g}^\phi > 1$ *and* f *is a nonzero element of* $\mathfrak{g}^{-\phi}$ *if* $\dim \mathfrak{g}^\phi = 1$. *In this case,* $f^l \cdot x_0$ *is a restricted weight vector in* X^ν *with restricted weight* $s_\phi \nu - \rho$.

Lemma 6.11 also gives some interesting information about the conical space of X^ν even when $\dim \mathfrak{a}$ is arbitrary. To see this, we need some general facts.

Lemma 6.13. *Let* $\Pi \subset \Sigma_+$ *be the set of simple restricted roots. Then the subalgebra* $\mathfrak{n}$ *of* $\mathfrak{g}$ *is generated by the subspaces* $\mathfrak{g}^\alpha$ *as* α *ranges through* Π.

Proof. We may, and do, assume that k is algebraically closed. For all $\psi \in \Sigma$, define the *order* $o(\psi)$ of ψ to be the integer $\Sigma\, n_\alpha$ $(\alpha \in \Pi)$, where the integers n_α are defined by the condition $\psi = \Sigma\, n_\alpha \alpha (\alpha \in \Pi)$. Then $\psi \in \Sigma_+$ if and only if $o(\psi) > 0$, and $\psi \in \Pi$ if and only if $o(\psi) = 1$. We shall show by induction on $o(\psi)$ $(\psi \in \Sigma_+)$ that $\mathfrak{g}^\psi$ lies in the space generated by the $\mathfrak{g}^\alpha$ $(\alpha \in \Pi)$. This is clearly true if $o(\psi) = 1$, so assume it is true for $\sigma(\psi) = m$ $(m \geq 1)$, and let $\psi' \in \Sigma_+$ have order $m + 1$. Then the standard theory of root systems shows that there exists $\alpha \in \Pi$ such that the scalar product $(\psi', \alpha) > 0$, and hence $\psi = \psi' - \alpha$ is a positive restricted root of order m. Define the subspace V of $\mathfrak{g}$ by $V = \amalg_{j=-\infty}^{\infty} \mathfrak{g}^{\psi + n\alpha}$, and construct as in §2 (taking α for ϕ) a subalgebra $\mathfrak{u}_\alpha$ of $\mathfrak{g}$ spanned by h_α, e_α and f_α. Then V is a $\mathfrak{u}_\alpha$-submodule of $\mathfrak{g}$, and ad h_α has eigenvalue $\psi(h_\alpha) + 2n$ on the subspace $\mathfrak{g}^{\psi + n\alpha}$ of V; in particular, $\mathfrak{g}^{\psi + n\alpha}$ is exactly the $(\psi(h_\alpha) + 2n)$-eigenspace for ad h_α in V. But

$$\psi(h_\alpha) + 2 = (\psi + \alpha)(h_\alpha) = \psi'(h_\alpha) > 0,$$

and so the integer $\psi(h_\alpha) \geq -1$. Hence ad h_α has eigenvalue ≥ -1 on $\mathfrak{g}^\psi$, and so by the representation theory of the three-dimensional simple Lie algebra $\mathfrak{u}_\alpha$, we see that $[e_\alpha, \mathfrak{g}^\psi] = \mathfrak{g}^{\psi+\alpha} = \mathfrak{g}^{\psi'}$, and so $[\mathfrak{g}^\alpha, \mathfrak{g}^\psi] = \mathfrak{g}^{\psi'}$. In view of the induction hypothesis, we are finished. Q.E.D.

Remark. The above proof is of course similar to the proof of Lemma 4.2.

Lemma 6.14. *Let* $\alpha \in \Pi$ *(see Lemma* 6.13*). Then* $s_\alpha\rho - \rho = s_\alpha\rho_\alpha - \rho_\alpha$, *and* $\rho(h_\alpha) = \rho_\alpha(h_\alpha)$.

Proof. The first assertion is proved in [7(b), Lemma 4.16]. It follows that

$$\rho - \rho_\alpha = s_\alpha(\rho - \rho_\alpha) = (\rho - \rho_\alpha) - (\rho - \rho_\alpha)(h_\alpha)\alpha,$$

and so $(\rho - \rho_\alpha)(h_\alpha) = 0$. Q.E.D.

Lemma 6.15. $\mathfrak{N}^-$ *is a direct sum of restricted weight spaces (with respect to the natural action of* $\mathfrak{a}$ *on* $\mathfrak{G}$*) with restricted weights consisting of those elements of* $\mathfrak{a}^*$ *of the form* $-\Sigma\, n_\beta\beta$, *where* β *ranges through* Π *and* $n_\beta \in \mathbf{Z}_+$. *Let* $\alpha \in \Pi$, *and suppose* $y \in \mathfrak{N}^-$ *is a restricted weight vector with restricted weight of the form* $c\alpha$ $(c \in k)$. *Then* $y \in \mathfrak{N}_{-\alpha}$ *and* $c \in -\mathbf{Z}_+$.

Proof. Let $\Sigma_+^1 = \{\psi \in \Sigma_+ | \tfrac{1}{2}\psi \notin \Sigma_+\}$. Then $\mathfrak{n}^- = \amalg\, \mathfrak{n}_{-\psi}$ as ψ ranges through Σ_+^1. Let $\psi_1, \psi_2, \ldots, \psi_p$ be the elements of Σ_+^1. Then the multiplication map in $\mathfrak{G}$ induces a linear isomorphism

$$\mathfrak{N}^- \simeq \mathfrak{N}_{-\psi_1} \otimes \mathfrak{N}_{-\psi_2} \otimes \cdots \otimes \mathfrak{N}_{-\psi_p}.$$

The lemma now follows easily. Q.E.D.

Lemma 6.16. *Let* $\alpha \in \Pi$, $\nu \in \mathfrak{a}^*$ *and* x_0 *the canonical generator of the twisted induced module* X^ν. *The sum of the restricted weight spaces of* X^ν *with restricted weights of the form* $\nu - \rho + c\alpha$ $(c \in k)$ *is exactly* $\mathfrak{N}_{-\alpha} \cdot x_0$.

Proof. This is clear from Lemma 6.15 and the fact that the linear isomorphism $\omega: \mathfrak{N}^- \to X^\nu$ which takes y to $y \cdot x_0$ raises restricted weights by $\nu - \rho$; i.e., if $y \in \mathfrak{N}^-$ is a restricted weight vector with restricted weight $\mu \in \mathfrak{a}^*$, then $\omega(y)$ is a restricted weight vector with restricted weight $\nu - \rho + \mu$. Q.E.D.

We now have the following generalization of Theorem 6.12:

Theorem 6.17. *Let* α *be a simple restricted root, and suppose* $2\alpha \notin \Sigma$. *Let* $\nu \in \mathfrak{a}^*$, *and let* Y *be the subspace of the twisted induced* $\mathfrak{g}$*-module* X^ν *spanned by the conical restricted weight vectors with restricted weights of*

the form $\nu - \rho + c\alpha$ $(c \in k)$. *Then* Y *is either one- or two-dimensional. Define* $h'_\alpha \in \mathfrak{a}$ *to be* h_α *if* $\dim \mathfrak{g}^\alpha > 1$ *and* $2h_\alpha$ *if* $\dim \mathfrak{g}^\alpha = 1$, *and let* x_0 *be the canonical generator of* X^ν. *If* $\nu(h'_\alpha)$ *is not a positive even integer, then* Y *is the span of* x_0. *Suppose* $\nu(h'_\alpha) = 2l$, l *a positive integer. Then* $\dim Y = 2$, *and* Y *has basis* $\{x_0, f^l \cdot x_0\}$, *where* $f = q_{-\alpha}$ *if* $\dim \mathfrak{g}^\alpha > 1$ *and* f *is a nonzero element of* $\mathfrak{g}^{-\alpha}$ *if* $\dim \mathfrak{g}^\alpha = 1$. *In this case,* $f^l \cdot x_0$ *is a restricted weight vector in* X^ν *with restricted weight* $s_\alpha \nu - \rho$.

Proof. Since the conical space of X^ν is clearly $\mathfrak{a}$-stable and hence the direct sum of its intersections with the restricted weight spaces of X^ν, $Y = (\mathfrak{N}_{-\alpha} \cdot x_0)^{\mathfrak{m}\oplus\mathfrak{n}}$ by Lemma 6.16. Let $y \in (\mathfrak{N}_{-\alpha} \cdot x_0)^{\mathfrak{m}\oplus\mathfrak{n}_\alpha}$, so that $y = u \cdot x_0$, where $u \in \mathfrak{N}_{-\alpha}$. Let β be a simple restricted root not equal to α. Then $\beta - \alpha$ is not a restricted root and is not zero, so that $[\mathfrak{g}^\beta, \mathfrak{n}_{-\alpha}] = [\mathfrak{g}^\beta, \mathfrak{g}^{-\alpha}] = 0$. Hence $[\mathfrak{g}^\beta, u] = 0$ in $\mathfrak{G}$, and so

$$\mathfrak{g}^\beta \cdot (u \cdot x_0) = u \cdot (\mathfrak{g}^\beta \cdot x_0) = 0.$$

Lemma 6.13 now shows that $y \in Y$. Thus $Y = (\mathfrak{N}_{-\alpha} \cdot x_0)^{\mathfrak{m}\oplus\mathfrak{n}_\alpha}$, and the theorem now follows from Lemmas 6.11 and 6.14. Q.E.D.

Remark. In the notation of Theorem 6.17, the assertion that $\nu(h'_\alpha)$ be a nonnegative even integer (possibly zero) is equivalent to the existence in X^ν of an $\mathfrak{m}$-invariant restricted weight vector with restricted weight $s_\alpha\nu - \rho = \nu - \rho - \nu(h_\alpha)\alpha$ (use Lemma 6.16). In this case, the $\mathfrak{m}$-invariant restricted weight vectors with restricted weight $s_\alpha\nu - \rho$ span a one-dimensional space and are conical vectors. Note also that if $\nu \in \mathfrak{a}^*$ is arbitrary and if f and x_0 are defined as in Theorem 6.17, then $f^m \cdot x_0$ (m a positive integer) is $\mathfrak{n}$-invariant if and only if its restricted weight is $s_\alpha\nu - \rho$.

We can reformulate our conclusions as follows:

Theorem 6.18. *Let* α *be a simple restricted root such that* $2\alpha \notin \Sigma$. *Let* $\mu, \nu \in \mathfrak{a}^*$, *and suppose that* $\mu - \nu$ *is of the form* $c\alpha$ $(c \in k)$. (*If* $\dim \mathfrak{a} = 1$, *then this is automatic.*) *Then* $\mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu)$ *is at most one-dimensional, and* $\dim \mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu) = 1$ *if and only if either* $\mu = \nu$, *or else* $\mu = s_\alpha\nu$ *and* $\nu(h'_\alpha)$ *is a nonnegative even integer, where* $h'_\alpha = h_\alpha$ *if* $\dim \mathfrak{g}^\alpha > 1$ *and* $h'_\alpha = 2h_\alpha$ *if* $\dim \mathfrak{g}^\alpha = 1$. *Also,* $\dim \mathrm{Hom}_{\mathfrak{g}}(x^\mu, X^\nu) = 1$ *if and only if* X^μ *is isomorphic to a* $\mathfrak{g}$*-submodule of* X^ν.

Proof. Recall from §2 that $\mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu)$ is isomorphic to the intersection Z of the conical space of X^ν with the restricted weight space for $\mu - \rho$. If $\mu = \nu$, then clearly $\dim Z = 1$. Suppose $\mu = s_\alpha\nu$ and $\nu(h'_\alpha)$ is a nonnegative even integer. Then the above remark implies that $\dim Z = 1$. Conversely, suppose $Z \neq 0$, so that X^ν contains a conical restricted weight vector x

with restricted weight $\mu-\rho$. Since $\mu=\nu+c\alpha$, $\mu-\rho=\nu-\rho+c\alpha$, and so $x\in Y$, in the notation of Theorem 6.17. If $\mu\neq\nu$, then x is not a multiple of x_0 (again in the notation of Theorem 6.17), so that $\nu(h'_\alpha)$ is a positive even integer and $\mu-\rho=s_\alpha\nu-\rho$, i.e., $\mu=s_\alpha\nu$, by Theorem 6.17. The last assertion of the theorem follows from the fact that any nonzero $\mathfrak{g}$-module map from X^μ into X^ν is injective (see §2). Q.E.D.

7. **The fundamental commutation relation in $\mathfrak{N}_{-\phi}$.** We shall continue to use the notation of §6, with k algebraically closed. But in this section, we explicitly assume that $\mathfrak{g}^{2\phi}\neq 0$, i.e., that $2\phi\in\Sigma$. We have the canonical elements $p_{-2\phi}\in S^2(\mathfrak{g}^{-2\phi})^{\mathfrak{m}}$ and $q_{-2\phi}=\lambda(p_{-2\phi})/2(\phi,\phi)\in\mathfrak{N}^{\mathfrak{m}}_{-\phi}$ (see §5).

It is clearly important to compute the commutator $[e_\phi, q_{-2\phi}]$ in $\mathfrak{G}$. This will easily turn out to be essentially $[f_\phi, q_{-\phi}]$, and we have to know to what extent this element commutes with $q_{-\phi}$. In particular, we want to compute $[[f_\phi, q_{-\phi}], q_{-\phi}]$. Lemma 6.4 also points out the importance of this commutator, since we need it in principle to simplify the commutator $[e_\phi, q^d_{-\phi}]$. It will turn out that $[[f_\phi, q_{-\phi}], q_{-\phi}]$ is essentially $f_\phi q_{-2\phi}$, and this is what we call the fundamental commutation relation in $\mathfrak{N}_{-\phi}$, the main result of this section. Because of this, we know how to compute the further commutators $[\cdots[[f_\phi, q_{-\phi}], q_{-\phi}]\cdots q_{-\phi}]$. The abstract algebraic setting in the next section will reveal a more precise reason for calling our relation "fundamental". The point will be that the fundamental relation and the trivial relation $f_\phi q_{-2\phi}=q_{-2\phi}f_\phi$ are in a sense *all* the relations involving f_ϕ, $q_{-\phi}$ and $q_{-2\phi}$.

Lemma 7.1. *The map* ad f_ϕ: $\mathfrak{g}_2^{-1}\to\mathfrak{g}_2^{-2}$ *is an isometry from* $4B_\theta|\mathfrak{g}_2^{-1}\times\mathfrak{g}_2^{-1}$ *to* $B_\theta|\mathfrak{g}_2^{-2}\times\mathfrak{g}_2^{-2}$.

Proof. Let $x, y\in\mathfrak{g}_2^{-1}$. Then

$$B_\theta([f_\phi,x],[f_\phi,y])=-B([f_\phi,x],\theta[f_\phi,y])=B([f_\phi,x],[e_\phi,\theta y])$$

$$=-B([e_\phi,[f_\phi,x]],\theta y)=-4B(x,\theta y)$$

(by Lemma 4.15)

$$=4B_\theta(x,y).\quad \text{Q.E.D.}$$

Recall from §6 the B_θ-orthogonal basis $\{f_1,\ldots,f_n\}$ of $\mathfrak{g}^{-\phi}$.

Lemma 7.2. *We have*

$$q_{-2\phi}=\frac{1}{16}\sum_{i=1}^{n}[f_\phi,f_i]^2=\frac{1}{16}\sum_{i=r+1}^{n}[f_\phi,f_i]^2.$$

Proof. By Lemma 7.1, $\{[f_\phi,f_{r+1}],\ldots,[f_\phi,f_n]\}$ is a B_θ-orthogonal

basis of $\mathfrak{g}_2^{-2} = \mathfrak{g}^{-2\phi}$ such that each $B_\theta([f_\phi, f_i], [f_\phi, f_i]) = 8/(\phi, \phi)$. Since $q_{-2\phi}$ is $1/2(\phi, \phi)$ times the sum of the squares of the elements of any B_θ-orthonormal basis of $\mathfrak{g}^{-2\phi}$, we must have $q_{-2\phi} = (1/16) \sum_{i=r+1}^n [f_\phi, f_i]^2$. But $[f_\phi, f_j] = 0$ if $j = 1, \dots, r$ and so the lemma follows. Q.E.D.

Lemma 7.3. *We have*

$$[e_\phi, q_{-2\phi}] = \frac{1}{4}[f_\phi, q_{-\phi}] = \frac{1}{2} \sum_{i=1}^n f_i [f_\phi, f_i]$$

$$= \frac{1}{2} \sum_{i=r+1}^n f_i [f_\phi, f_i].$$

Proof. By Lemma 7.2,

$$[e_\phi, q_{-2\phi}] = \frac{1}{16} \sum_{i=r+1}^n [e_\phi, [f_\phi, f_i]^2]$$

$$= \frac{1}{16} \sum_{i=r+1}^n ([e_\phi, [f_\phi, f_i]][f_\phi, f_i]$$

$$+ [f_\phi, f_i][e_\phi, [f_\phi, f_i]])$$

$$= \frac{1}{4} \sum_{i=r+1}^n (f_i [f_\phi, f_i] + [f_\phi, f_i] f_i)$$

(by Lemma 4.15)

$$= \frac{1}{2} \sum_{i=r+1}^n f_i [f_\phi, f_i]$$

(since $[f_\phi, f_i] \in \mathfrak{g}^{-2\phi}$, which is central in $\mathfrak{N}_{-\phi}$)

$$= \frac{1}{2} \sum_{i=1}^n f_i [f_\phi, f_i].$$

On the other hand, $q_{-\phi} = \sum_{i=1}^n f_i^2$, so that

$$[f_\phi, q_{-\phi}] = \sum_{i=1}^n [f_\phi, f_i^2]$$

$$= \sum_{i=1}^n ([f_\phi, f_i] f_i + f_i [f_\phi, f_i])$$

$$= 2 \sum_{i=1}^n f_i [f_\phi, f_i]. \quad \text{Q.E.D.}$$

Theorem 7.4. (*The fundamental commutation relation in* $\mathfrak{N}_{-\phi}$.) *We have*

$$[[f_\phi, q_{-\phi}], q_{-\phi}] = -64 f_\phi q_{-2\phi}.$$

More generally, suppose the field k is arbitrary of characteristic zero, and let $f \in \mathfrak{g}^{-\phi}$. Then

$$[[f, q_{-\phi}], q_{-\phi}] = -64 f q_{-2\phi}.$$

Proof. It is clearly sufficient to prove the first assertion. But by Lemma 7.3,

$$[[f_\phi, q_{-\phi}], q_{-\phi}] = 4[[e_\phi, q_{-2\phi}], q_{-\phi}]$$
$$= 4[[e_\phi, q_{-\phi}], q_{-2\phi}] = 8 f_\phi [h_\phi, q_{-2\phi}],$$

by Lemma 6.4, and this is just $-64 f_\phi q_{-2\phi}$. Q.E.D.

8. The transfer principles. Here we assume that $\mathfrak{g}^{2\phi} \neq 0$, as in §7. But we take k to be an arbitrary field of characteristic zero.

If we attempt to compute directly the conical vectors in the twisted induced modules X^ν ($\nu \in \mathfrak{a}^*$), we are confronted with monumental difficulties (cf. the remark at the end of this section). Trying to avoid these problems, we discovered a metamathematical "transfer principle" (Theorem 8.6) which enables us essentially to transfer certain theorems about conical vectors in modules over one semisimple symmetric Lie algebra to theorems about conical vectors in modules over any other semisimple symmetric Lie algebra. This reduces the problem of computing certain conical vectors to any one special case of semisimple symmetric Lie algebra (in which twice the relevant simple restricted root is a restricted root). The proof of this "transfer principle for conical vectors" is based on another metamathematical result (Theorem 8.4) which states that certain kinds of algebraic identities in $\mathfrak{N}_{-\phi}$ can be transferred from one semisimple symmetric Lie algebra to another. The starting point for the proof of this theorem is the "fundamental commutation relation" of the last section.

Let $P = k[w, x, y, z]$, the polynomial algebra in four indeterminates, and define a P-module structure on $\mathfrak{N}_{-\phi}$ by the correspondences

$w \mapsto$ left multiplication by $q_{-\phi}$,

$x \mapsto$ left multiplication by $q_{-2\phi}$,

$y \mapsto$ right multiplication by $q_{-\phi}$,

$z \mapsto$ right multiplication by $q_{-2\phi}$.

This P-module structure is well defined because $[q_{-\phi}, q_{-2\phi}] = 0$ in $\mathfrak{N}_{-\phi}$.

Theorem 8.1. *Let f be an arbitrary B_θ-nonisotropic element of $\mathfrak{g}^{-\phi}$, and let P^f denote the annihilator of f in P under the above module action.*

Then the ideal P^f *is generated by* $x - z$ *and* $w^2 - 2wy + y^2 + 64x$, *that is,*

$$P^f = P(x-z) + P(w^2 - 2wy + y^2 + 64x).$$

Proof. Since $q_{-2\phi}$ is central in $\mathfrak{N}_{-\phi}$, it is clear that $x - z \in P^f$, and so $P(x-z) \subset P^f$. The fundamental commutation relation, Theorem 7.4, implies immediately that $w^2 - 2wy + y^2 + 64x \in P^f$, and hence the ideal generated by this element is contained in P^f. What we must show now is that these two ideals generate P^f.

Let $a \in P^f$, $a \neq 0$, and regard P as $k[x, y, z][w]$. Since the leading coefficient 1 of $w^2 - 2wy + y^2 + 64x$ is a unit in $k[x, y, z]$, the Euclidean algorithm implies the existence of $s, t \in k[x, y, z][w]$, where t is a polynomial of degree at most 1 in w, such that

$$a = s(w^2 - 2wy + y^2 + 64x) + t.$$

Here t is of the form $u + wv$, where $u, v \in k[x, y, z]$. Since $a \in P^f$, $t \in P^f$. Also, there exist polynomials u', $v' \in k[y, z]$ such that

$$u \equiv u' \pmod{P(x-z)} \quad \text{and} \quad v \equiv v' \pmod{P(x-z)}.$$

Hence

$$t \equiv u' + wv' \pmod{P(x-z)},$$

and so

$$a \equiv u' + wv' \pmod{P(x-z) + P(w^2 - 2wy + y^2 + 64x)}.$$

In particular, $u' + wv' \in P^f$. Write $u' = u'(y, z)$ and $v' = v'(y, z)$. Then by the definition of the module action of P on $\mathfrak{N}_{-\phi}$, we have

$$fu'(q_{-\phi}, q_{-2\phi}) + q_{-\phi} f v'(q_{-\phi}, q_{-2\phi}) = 0,$$

and so

$$f(u'(q_{-\phi}, q_{-2\phi}) + q_{-\phi} v'(q_{-\phi}, q_{-2\phi})) - [f, q_{-\phi}] v'(q_{-\phi}, q_{-2\phi}) = 0.$$

Set $\alpha(y, z) = u'(y, z) + yv'(y, z)$ and $\beta(y, z) = -v'(y, z)$ ($\alpha, \beta \in k[y, z]$). Then

$$f\alpha(q_{-\phi}, q_{-2\phi}) + [f, q_{-\phi}]\beta(q_{-\phi}, q_{-2\phi}) = 0.$$

It is sufficient to show that $\alpha = \beta = 0$, since then we will have $u' = v' = 0$, and so $a \in P(x-z) + P(w^2 - 2wy + y^2 + 64x)$.

As in the proof of Theorem 5.1, let $\mathfrak{N}_0 \subset \mathfrak{N}_1 \subset \mathfrak{N}_2 \subset \cdots$ be the usual filtration of $\mathfrak{N}_{-\phi}$, and for each $r \in \mathbf{Z}_+$, let $\pi_r : \mathfrak{N}_r \to \mathfrak{N}_r/\mathfrak{N}_{r-1}$ be the canonical map. (Here $\mathfrak{N}_{-1} = 0$.) Also, let $\sigma_r : S^r(\mathfrak{n}_{-\phi}) \to \mathfrak{N}_r/\mathfrak{N}_{r-1}$ be the natural map,

so that σ_r is a linear isomorphism by the Poincaré-Birkhoff-Witt theorem.

Write

$$\alpha(y, z) = \sum_{j=0}^{c} \sum_{i=0}^{j} a_{ij} y^i z^{j-i}$$

and

$$\beta(y, z) = \sum_{j=0}^{d} \sum_{i=0}^{j} b_{ij} y^i z^{j-i}$$

with $c, d \in \mathbf{Z}_+$ and $a_{ij}, b_{ij} \in k$. If $\alpha \neq 0$, we may assume that some $a_{ic} \neq 0$ $(i = 0, \ldots, c)$, and if $\beta \neq 0$, we may also assume that some $b_{id} \neq 0$ $(i = 0, \ldots, d)$. Also, if $\alpha = 0$, take $c = 0$ and if $\beta = 0$, take $d = 0$.

Now we claim that $[f, q_{-\phi}] \in \mathfrak{N}_2$ and $[f, q_{-\phi}] \notin \mathfrak{N}_1$. In fact, it is sufficient to prove this when k is algebraically closed. But then a suitable multiple of f may be taken as the f_ϕ of §7, and the claim follows from Lemma 7.3. In particular, $f\alpha(q_{-\phi}, q_{-2\phi}) \in \mathfrak{N}_{2c+1}$ and $[f, q_{-\phi}]\beta(q_{-\phi}, q_{-2\phi}) \in \mathfrak{N}_{2d+2}$; recall that the sum of these two terms is zero. Either $2c + 1 > 2d + 2$ or $2c + 1 < 2d + 2$. Suppose the first inequality holds. Then

$$\pi_{2c+1}(f\alpha(q_{-\phi}, q_{-2\phi})) = 0,$$

so that

$$\pi_{2c+1}\left(f \sum_{i=0}^{c} a_{ic} q_{-\phi}^i q_{-2\phi}^{c-i}\right) = 0.$$

Let

$$p'_{-\phi} = \frac{2}{(\phi, \phi)} p_{-\phi} \quad \text{and} \quad p'_{-2\phi} = \frac{1}{2(\phi, \phi)} p_{-2\phi},$$

and set

$$s = f \sum_{i=0}^{c} a_{ic} (p'_{-\phi})^i (p'_{-2\phi})^{c-i} \in S^{2c+1}(\mathfrak{n}_{-\phi}).$$

Then

$$\sigma_{2c+1}(s) = \pi_{2c+1}(\lambda(s)) = \pi_{2c+1}\left(\lambda(f) \sum_{i=0}^{c} a_{ic} \lambda(p'_{-\phi})^i \lambda(p'_{-2\phi})^{c-i}\right)$$

$$= \pi_{2c+1}\left(f \sum_{i=0}^{c} a_{ic} q_{-\phi}^i q_{-2\phi}^{c-i}\right) = 0.$$

Hence $s = 0$, and so each $a_{ic} = 0$ $(i = 0, \ldots, c)$. This is only possible if $\alpha = 0$. But then $c = 0$, and the inequality $2c + 1 > 2d + 2$ cannot hold. Hence we may assume that $2c + 1 < 2d + 2$. In this case,

$$\pi_{2d+2}([f, q_{-\phi}]\beta(q_{-\phi}, q_{-2\phi})) = 0,$$

and so

$$\pi_{2d+2}\left([f, q_{-\phi}] \sum_{i=0}^{d} b_{id} q_{-\phi}^{i} q_{-2\phi}^{d-i}\right) = 0.$$

Since $[f, q_{-\phi}] \notin \mathfrak{N}_1$ (see above), there exists a nonzero element $g \in S^2(\mathfrak{n}_{-\phi})$ such that $\lambda(g) \equiv [f, q_{-\phi}] \pmod{\mathfrak{N}_1}$. Set

$$h = g \sum_{i=0}^{d} b_{id}(p'_{-\phi})^{i}(p'_{-2\phi})^{d-i} \in S^{2d+2}(\mathfrak{n}_{-\phi}).$$

Then

$$\sigma_{2d+2}(h) = \pi_{2d+2}(\lambda(h)) = \pi_{2d+2}\left(\lambda(g) \sum_{i=0}^{d} b_{id}\lambda(p'_{-\phi})^{i}\lambda(p'_{-2\phi})^{d-i}\right)$$

$$= \pi_{2d+2}\left([f, q_{-\phi}] \sum_{i=0}^{d} b_{id} q_{-\phi}^{i} q_{-2\phi}^{d-i}\right) = 0.$$

Hence $h = 0$. But $g \neq 0$, so that each $b_{id} = 0$ $(i = 0, \ldots, d)$. This proves that $\beta = 0$, and so $d = 0$. Since $2c + 1 < 2d + 2$, we also have $c = 0$. Thus α is a scalar, and the equation $f\alpha = 0$ shows that $\alpha = 0$. We have proved that $\alpha = \beta = 0$, and hence the theorem. Q.E.D.

Suppose now that $\dim \mathfrak{g}^{2\phi} = 1$, and suppose there exists an element $r_{-2\phi} \in \mathfrak{g}^{-2\phi}$ such that $r_{-2\phi}^2 = q_{-2\phi}$ in $\mathfrak{N}_{-\phi}$. (Such an element exists if k is algebraically closed, but otherwise, it might not exist.) Define a new P-module structure on $\mathfrak{N}_{-\phi}$ by the correspondences

$w \mapsto$ left multiplication by $q_{-\phi}$,

$x \mapsto$ left multiplication by $r_{-2\phi}$,

$y \mapsto$ right multiplication by $q_{-\phi}$,

$z \mapsto$ right multiplication by $r_{-2\phi}$.

This P-module structure is well defined since $[q_{-\phi}, r_{-2\phi}] = 0$ in $\mathfrak{N}_{-\phi}$.

Theorem 8.2. *Under the above hypotheses, let $f \in \mathfrak{g}^{-\phi}$ be B_θ-nonisotropic, and let P_f be the annihilator of f in P under the new module action. Then*

$$P_f = P(x - z) + P(w^2 - 2wy + y^2 + 64x^2).$$

Proof. The first part of the proof of Theorem 8.1 carries over to the present situation and shows that is sufficient to prove the following: Let

$\alpha(y, z)$, $\beta(y, z) \in k[y, z]$, and suppose

$$f\alpha(q_{-\phi}, r_{-2\phi}) + [f, q_{-\phi}]\beta(q_{-\phi}, r_{-2\phi}) = 0.$$

Then $\alpha = \beta = 0$.

It is clearly sufficient to assume that k is algebraically closed and that f is the element f_ϕ of §§6 and 7. But then by Lemma 7.3, $[f_\phi, q_{-\phi}] = 2f_n[f_\phi, f_n]$ (where f_n is as in that lemma; see §6), since $\dim \mathfrak{g}^{2\phi} = 1$. By Lemma 7.2, $[f_\phi, f_n]^2 = 16q_{-2\phi}$ in $\mathfrak{N}_{-\phi}$, and since $[f_\phi, f_n] \in \mathfrak{g}^{-2\phi}$, we must have $4r_{-2\phi} = \pm[f_\phi, f_n]$. Changing the sign of f_n if necessary, we may assume that $4r_{-2\phi} = [f_\phi, f_n]$. Setting $\alpha'(y, z) = \alpha(y, z)$ and $\beta'(y, z) = 8z\beta(y, z)$ in $k[y, z]$, we have

$$f_\phi\alpha'(q_{-\phi}, r_{-2\phi}) + f_n\beta'(q_{-\phi}, r_{-2\phi}) = 0, \tag{2}$$

and it is sufficient to show that $\alpha' = \beta' = 0$.

Now $[e_\phi, f_n] \in \mathfrak{m}$ (where e_ϕ is as in §6), by Lemma 6.2, and so

$$[e_\phi, f_n] \cdot \alpha'(q_{-\phi}, r_{-2\phi}) = [e_\phi, f_n] \cdot \beta'(q_{-\phi}, r_{-2\phi}) = 0$$

in $\mathfrak{N}_{-\phi}$, since $q_{-\phi}, r_{-2\phi} \in \mathfrak{N}^{\mathfrak{m}}_{-\phi}$. Also, $[[e_\phi, f_n], f_\phi] = -6f_n$ by Lemma 4.15 and $[[e_\phi, f_n], f_n] = 6f_\phi$ by Lemma 6.3. Hence the application of $[e_\phi, f_n]$ to (*) gives

$$f_n\alpha'(q_{-\phi}, r_{-2\phi}) - f_\phi\beta'(q_{-\phi}, r_{-2\phi}) = 0. \tag{3}$$

Abbreviate $\alpha'(q_{-\phi}, r_{-2\phi})$ by α_0 and $\beta'(q_{-\phi}, r_{-2\phi})$ by β_0. Multiplying (2) on the right by α_0, multiplying (3) on the right by $-\beta_0$, and adding the two results, we get $f_\phi(\alpha_0^2 + \beta_0^2) = 0$. Since $\mathfrak{G}$ has no zero divisors,

$$(\alpha_0 + (-1)^{1/2}\beta_0)(\alpha_0 - (-1)^{1/2}\beta_0) = \alpha_0^2 + \beta_0^2 = 0,$$

and so $\alpha_0 = \pm(-1)^{1/2}\beta_0$. Thus (*) implies that $\alpha_0 = \beta_0 = 0$. The fact that $\alpha'(y, z) = \beta'(y, z) = 0$ now follows from Theorem 5.1, Case 3. Q.E.D.

Now assume the original hypotheses of this section, so that $\mathfrak{g}^{2\phi} \neq 0$. The following consequence of the last two theorems is immediate:

Corollary 8.3. *Let Q be the polynomial algebra in two variables over k, and let $a_i, b_i \in Q$ ($i = 1, \ldots, r$, $r \in \mathbf{Z}_+$). Let f be a B_θ-nonisotropic element of $\mathfrak{g}^{-\phi}$. Then*

$$\sum_{i=1}^{r} a_i(q_{-\phi}, q_{-2\phi})fb_i(q_{-\phi}, q_{-2\phi}) = 0 \tag{4}$$

in $\mathfrak{N}_{-\phi}$ if and only if

$$\sum_{i=1}^{r} a_i \otimes b_i \in P(x-z) + P(w^2 - 2wy + y^2 + 64x),$$

where we identify P *with* $Q \otimes Q$ *in the natural way. Suppose in addition that* $\dim \mathfrak{g}^{2\phi} = 1$ *and that there exists an element* $r_{-2\phi} \in \mathfrak{g}^{-2\phi}$ *such that* $r_{-2\phi}^2 = q_{-2\phi}$. *Then*

$$\sum_{i=1}^{r} a_i(q_{-\phi}, r_{-2\phi}) f b_i(q_{-\phi}, r_{-2\phi}) = 0 \tag{5}$$

in $\mathfrak{N}_{-\phi}$ *if and only if*

$$\sum_{i=1}^{r} a_i \otimes b_i \in P(x-z) + P(w^2 - 2wy + y^2 + 64x^2),$$

where we again identify P *with* $Q \otimes Q$.

This corollary proves:

Theorem 8.4. (*The transfer principle for* $\mathfrak{N}_{-\phi}$.) *Let* Q *be the polynomial algebra in two variables over* k, *and let* $a_i, b_i \in Q$ $(i = 1, \ldots, r, r \in \mathbb{Z}_+)$. *Let* $(\mathfrak{g}, \theta)$ *be a semisimple symmetric Lie algebra over* k *with symmetric decomposition* $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$, $\mathfrak{a}$ *a splitting Cartan subspace of* $\mathfrak{p}$, $\Sigma \subset \mathfrak{a}^*$ *the corresponding system of restricted roots,* $\phi \in \Sigma$ *such that* $2\phi \in \Sigma$, $\mathfrak{N}_{-\phi}$ *the universal enveloping algebra of the Lie subalgebra* $\mathfrak{n}_{-\phi} = \mathfrak{g}^{-\phi} \oplus \mathfrak{g}^{-2\phi}$ *of* $\mathfrak{g}$, λ: $S(\mathfrak{n}_{-\phi}) \to \mathfrak{N}_{-\phi}$ *the canonical linear isomorphism,* B *the Killing form of* $\mathfrak{g}$, B_θ *the symmetric bilinear form on* $\mathfrak{g}$ *defined by the condition* $B_\theta(x, y) = -B(x, \theta y)$ *for all* $x, y \in \mathfrak{g}$, f *a* B_θ*-nonisotropic vector in* $\mathfrak{g}^{-\phi}$, $p_{-\phi} \in S^2(\mathfrak{g}^{-\phi})$ *and* $p_{-2\phi} \in S^2(\mathfrak{g}^{-2\phi})$ *the canonical elements defined by* B_θ, *and* $q_{-\phi} = 2\lambda(p_{-\phi})/(\phi, \phi)$ *and* $q_{-2\phi} = \lambda(p_{-2\phi})/2(\phi, \phi) \in \mathfrak{N}_{-\phi}$. *Then the truth or falsity of equation* (4) *in* $\mathfrak{N}_{-\phi}$ *depends only on* a_i *and* b_i $(i = 1, \ldots, r)$ *and not on* $\mathfrak{g}, \theta, \mathfrak{a}, \phi$ *or* f. *Moreover, suppose in addition that* $\dim \mathfrak{g}^{2\phi} = 1$ *and that there exists an element* $r_{-2\phi} \in \mathfrak{g}^{-2\phi}$ *such that* $r_{-2\phi}^2 = q_{-2\phi}$. *Then the truth or falsity of equation* (5) *in* $\mathfrak{N}_{-\phi}$ *depends only on* a_i *and* b_i $(i = 1, \ldots, r)$, *and not on* $\mathfrak{g}, \theta, \mathfrak{a}, \phi, f$ *or* $r_{-2\phi}$.

In order to apply this theorem to conical vectors, we need:

Lemma 8.5. *Suppose* ϕ *and* $2\phi \in \Sigma_+$, *and let* V *be a* $\mathfrak{g}$*-module,* $v \in V^{\mathfrak{m} \oplus \mathfrak{n}_\phi}$ *a restricted weight vector with restricted weight* $\mu \in \mathfrak{a}^*$, $e_0 \in \mathfrak{g}^\phi$ *and* $i, j \in \mathbb{Z}_+$. *Then*

$$e_0 \cdot (q_{-2\phi}^j q_{-\phi}^i \cdot v) = y_{ij} \cdot v,$$

where $y_{ij} \in \mathfrak{N}_{-\phi}$ *is given by the formula*

$$y_{ij} = -\frac{1}{4}j[\theta e_0, q_{-\phi}]q_{-2\phi}^{j-1}q_{-\phi}^i$$

$$-\sum_{m=1}^{i} 2((\mu + \rho_\phi)(h_\phi) + 2 - 4m)q_{-2\phi}^j q_{-\phi}^{i-m}(\theta e_0)q_{-\phi}^{m-1},$$

where ρ_ϕ *is as in Lemma 6.4. Moreover, suppose in addition that* $\dim \mathfrak{g}^{2\phi} = 1$ *and that there exists an element* $r_{-2\phi} \in \mathfrak{g}^{-2\phi}$ *such that* $r_{-2\phi}^2 = q_{-2\phi}$. *Then*

$$r_{-2\phi}e_0 \cdot (r_{-2\phi}^j q_{-\phi}^i \cdot v) = y'_{ij} \cdot v,$$

where $y'_{ij} \in \mathfrak{N}_{-\phi}$ *is given by the formula*

$$y'_{ij} = -\frac{1}{8}j[\theta e_0, q_{-\phi}]r_{-2\phi}^{j-1}q_{-\phi}^i$$

$$-\sum_{m=1}^{i} 2((\mu + \rho_\phi)(h_\phi) + 2 - 4m)r_{-2\phi}^{j+1}q_{-\phi}^{i-m}(\theta e_0)q_{-\phi}^{m-1}.$$

Proof. We may assume that k is algebraically closed and that $e_0 = e_\phi$, so that $\theta e_0 = -f_\phi$. To prove the first assertion, note that

$$e_\phi \cdot (q_{-2\phi}^j q_{-\phi}^i \cdot v) = \sum_{l=1}^{j} q_{-2\phi}^{j-l}[e_\phi, q_{-2\phi}]q_{-2\phi}^{l-1}q_{-\phi}^i \cdot v$$

$$+ \sum_{m=1}^{i} q_{-2\phi}^j q_{-\phi}^{i-m}[e_\phi, q_{-\phi}]q_{-\phi}^{m-1} \cdot v.$$

By Lemma 7.3, the first term on the right is $\frac{1}{4}j[f_\phi, q_{-\phi}]q_{-2\phi}^{j-1}q_{-\phi}^i \cdot v$. To handle the second term, use Lemma 6.4. Since $v \in V^{\mathfrak{m}}$, Lemma 6.2 shows that the second term is

$$\sum_{m=1}^{i} 2q_{-2\phi}^j q_{-\phi}^{i-m}((\rho_\phi - \phi)(h_\phi)f_\phi + f_\phi h_\phi)q_{-\phi}^{m-1} \cdot v.$$

But it was shown in the proof of Lemma 6.5 that $h_\phi q_{-\phi}^{m-1} = q_{-\phi}^{m-1}(h_\phi - 4(m-1))$. Thus the term becomes

$$\sum_{m=1}^{i} 2((\mu + \rho_\phi)(h_\phi) + 2 - 4m)q_{-2\phi}^j q_{-\phi}^{i-m} f_\phi q_{-\phi}^{m-1} \cdot v,$$

and this proves the first assertion of the lemma.

Now suppose that $\dim \mathfrak{g}^{2\phi} = 1$ and that $r_{-2\phi}^2 = q_{-2\phi}$ ($r_{-2\phi} \in \mathfrak{g}^{-2\phi}$). Then

$$r_{-2\phi}e_\phi \cdot (r^j_{-2\phi}q^i_{-\phi} \cdot v) = \sum_{l=1}^{j} r^{j-l+1}_{-2\phi}[e_\phi, r_{-2\phi}]r^{l-1}_{-2\phi}q^i_{-\phi} \cdot v$$

$$+ \sum_{m=1}^{i} r^{j+1}_{-2\phi}q^{i-m}_{-\phi}[e_\phi, q_{-\phi}]q^{m-1}_{-\phi} \cdot v.$$

The second term is treated exactly as in the first part of the proof, and all that remains is to show that the first term is $(1/8)j[f_\phi, q_{-\phi}]r^{j-1}_{-2\phi}q^i_{-\phi} \cdot v$. But $[f_\phi, q_{-\phi}] = 2f_n[f_\phi, f_n]$ and $[f_\phi, f_n] = \pm 4r_{-2\phi}$ as in the proof of Theorem 8.2, and so

$$[f_\phi, q_{-\phi}] = \pm 8 f_n r_{-2\phi}$$

and

$$[e_\phi, r_{-2\phi}] = \pm\frac{1}{4}[e_\phi, [f_\phi, f_n]] = \pm f_n,$$

by Lemma 4.15. Thus the two indicated terms are equal, and the lemma is proved. Q.E.D.

We can now prove:

Theorem 8.6. (*The transfer principle for conical vectors.*) *Let Q be the polynomial algebra in two variables over k, and let $a_0 \in Q$. Also, let $c_0 \in k$. In continuation of the notation of Theorem 8.4, let Σ_+ be a positive system in Σ, $\alpha \in \Sigma_+$ a simple restricted root such that $2\alpha \in \Sigma$, $h_\alpha \in \mathfrak{a}$ as defined in §2, $\nu \in \mathfrak{a}^*$ such that $\nu(h_\alpha) = c_0$, X^ν the twisted induced $\mathfrak{g}$-module (see §2) and $x_0 \in X^\nu$ the canonical generator. Then the truth or falsity of the assertion "$a_0(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ is a conical vector in X^ν" depends only on a_0 and c_0, and not on $\mathfrak{g}$, θ, $\mathfrak{a}$, Σ_+, α or ν (except that $\nu(h_\alpha) = c_0$). Moreover, suppose in addition that $\dim \mathfrak{g}^{2\alpha} = 1$ and that there exists an element $r_{-2\alpha} \in \mathfrak{g}^{-2\alpha}$ such that $r^2_{-2\alpha} = q_{-2\alpha}$. Then the truth or falsity of the assertion "$a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ is a conical vector in X^ν" depends only on a_0 and c_0, and not on $\mathfrak{g}$, θ, $\mathfrak{a}$, Σ_+, α, $r_{-2\alpha}$ or ν (where $\nu(h_\alpha) = c_0$).*

Proof. Write $Q = k[x, y]$ and $a_0 = \Sigma^t_{i,j=0} b_{ij}x^iy^j$ ($t \in \mathbb{Z}_+$ and $b_{ij} \in k$) and assume $a_0 \neq 0$. In view of Theorem 5.1 (Cases 3 and 4), $a_0(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ is a nonzero $\mathfrak{m}$-invariant vector in X^ν. Let e_0 be a B_θ-nonisotropic vector in $\mathfrak{g}^\alpha$. Then by Corollary 4.3 and Lemma 6.13 (see the proof of Theorem 6.17), $e_0 \cdot (a_0(q_{-\alpha}, q_{-2\alpha}) \cdot x_0) = 0$ if and only if $a_0(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ is conical. But by Lemma 8.5, this is the case if and only if

$$\sum_{i,j=0}^{t} b_{ij}y_{ij} = 0 \quad \text{in } \mathfrak{N}_{-\alpha}, \tag{6}$$

where $y_{ij} \in \mathfrak{N}_{-\alpha}$ is as in Lemma 8.5, with ϕ replaced by α and μ by $\nu-\rho$ ($\rho = \frac{1}{2}\Sigma(\dim \mathfrak{g}^{\psi})\psi$, $\psi \in \Sigma_+$). But $(\nu - \rho + \rho_\alpha)(h_\alpha) = \nu(h_\alpha) = c_0$ by Lemma 6.14, so that

$$y_{ij} = -\frac{1}{4} j(\theta e_0) q_{-2\alpha}^{j-1} q_{-\alpha}^{i+1} + \frac{1}{4} j q_{-\alpha}(\theta e_0) q_{-2\alpha}^{j-1} q_{-\alpha}^{i}$$

$$- \sum_{m=1}^{i} 2(c_0 + 2 - 4m) q_{-2\alpha}^{j} q_{-\alpha}^{i-m} (\theta e_0) q_{-\alpha}^{m-1}.$$

Since θe_0 is a B_θ-nonisotropic vector in $\mathfrak{g}^{-\alpha}$, (6) is an equation of the form treated in Theorem 8.4, with ϕ replaced by α, and with the a_i and b_i in Theorem 8.4 dependent only on a_0 and c_0. That theorem now implies the first assertion of the present one.

Now assume that $\dim \mathfrak{g}^{2\alpha} = 1$ and that $r_{-2\alpha}^2 = q_{-2\alpha}$ ($r_{-2\alpha} \in \mathfrak{g}^{-2\alpha}$), and let $a_0 \neq 0$ and e_0 be as above. By Case 3 of Theorem 5.1, $a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ is a nonzero $\mathfrak{m}$-invariant vector in X^ν. Also, since $r_{-2\alpha}$ is a nonzero element of $\mathfrak{N}^-$, $r_{-2\alpha} e_0 \cdot (a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0) = 0$ if and only if

$$e_0 \cdot (a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0) = 0,$$

and this is true if and only if $a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ is conical, as above. Combining the last parts of Lemma 8.5 and Theorem 8.4 as above, we get the last assertion of the theorem. Q. E. D.

Remark. Of course, the above proof in principle provides an explicit reformulation of the assertion "$a_0(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ is a conical vector in X^ν" in terms of a_0 and c_0 alone, and similarly for $a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$, under the extra hypotheses. But these reformulations are much too complicated to be useful in determining directly the conical vectors in the induced modules X^ν. Instead, we shall compute the conical vectors for a special $\mathfrak{g}$ (see §9), and then use Theorem 8.6 to obtain them for general $\mathfrak{g}$. The determination of the conical vectors in the special case is not trivial, but at least it *can* be done.

9. A special case. Following the plan indicated by Theorem 8.6, we shall determine all the conical vectors in all the twisted induced modules X^ν ($\nu \in \mathfrak{a}^*$) for a special semisimple symmetric Lie algebra $(\mathfrak{g}, \theta)$. Here $(\mathfrak{g}, \theta)$ will have essentially the same structure as the real semisimple Lie algebra $\mathfrak{su}(2, 1)$. Our methods will be special; in fact, one of our main points is that it is too difficult to compute directly the conical vectors in general (cf. §8). We are grateful to L. Corwin and N. Wallach for their help in carrying out this special case (see the introduction).

Assume k is algebraically closed. Let $\mathfrak{g} = \mathfrak{sl}(3, k)$, the simple Lie al-

gebra of all traceless 3×3 matrices over k. Let $i = (-1)^{1/2}$, and let $\mathfrak{k} \subset \mathfrak{g}$ and $\mathfrak{p} \subset \mathfrak{g}$ be the spaces of matrices

$$\left\{\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & ia_{21} \\ -a_{13} & -ia_{12} & a_{11} \end{pmatrix}\right\} \quad \text{and} \quad \left\{\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & 0 & -ib_{21} \\ b_{13} & ib_{12} & -b_{11} \end{pmatrix}\right\},$$

respectively, where a_{ij}, $b_{ij} \in k$ and $2a_{11} + a_{22} = 0$. Then $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$, $[\mathfrak{k}, \mathfrak{k}] \subset \mathfrak{k}$, $[\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p}$ and $[\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k}$, so that the linear automorphism θ of $\mathfrak{g}$ which is 1 on $\mathfrak{k}$ and -1 on $\mathfrak{p}$ is a Lie algebra automorphism. Thus $(\mathfrak{g}, \theta)$ is a semisimple symmetric Lie algebra with symmetric decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$.

For all l, $m = 1, 2, 3$, let E_{lm} denote the 3×3 matrix which is 1 in the (l, m)-entry and 0 in all other entries. Let $\mathfrak{a}$ be the one-dimensional subspace of $\mathfrak{p}$ spanned by the matrix $h = 2(E_{11} - E_{33})$. Then $\mathfrak{a}$ is a splitting Cartan subspace of $\mathfrak{p}$. Let α be the linear functional on $\mathfrak{a}$ which is 2 on h. Then the set Σ of restricted roots of $\mathfrak{g}$ with respect to $\mathfrak{a}$ is $\{\pm\alpha, \pm 2\alpha\}$, $\mathfrak{g}^0$ is the set of traceless diagonal matrices, $\mathfrak{g}^{\alpha}$ is the span of E_{12} and E_{23}, $\mathfrak{g}^{-\alpha}$ is the span of E_{21} and E_{32}, $\mathfrak{g}^{2\alpha}$ is the span of E_{13}, and $\mathfrak{g}^{-2\alpha}$ is the span of E_{31}. Also, let h' be the matrix $E_{11} - 2E_{22} + E_{33}$. Then the centralizer $\mathfrak{m}$ of $\mathfrak{a}$ in $\mathfrak{k}$ is the span of h', and $\mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a}$.

Let Σ_+ be the positive system in Σ consisting of α and 2α. Then α is the unique simple restricted root. Since $\alpha(h) = 2$, $h = h_\alpha$ as defined in §2.

The Killing form B of $\mathfrak{g}$ is given by the formula $B(x, y) = 6 \operatorname{tr} xy$. Thus on $\mathfrak{g}^{-\alpha}$, the form $B_\theta(x, y) = -B(x, \theta y)$ is given by the formula

$$B_\theta(aE_{21} + bE_{32}, cE_{21} + dE_{32}) = -6i(ad + bc)$$

$(a, b, c, d \in k)$, and on $\mathfrak{g}^{-2\alpha}$, B_θ is given by

$$B_\theta(aE_{31}, bE_{31}) = 6ab$$

$(a, b \in k)$. Hence $\{(12)^{-1/2}(E_{21} + iE_{32}), (12)^{-1/2}(iE_{21} + E_{32})\}$ is a B_θ-orthonormal basis of $\mathfrak{g}^{-\alpha}$, and $\{6^{-1/2}E_{31}\}$ is a B_θ-orthonormal basis of $\mathfrak{g}^{-2\alpha}$. Since the canonical elements $p_{-\alpha} \in S^2(\mathfrak{g}^{-\alpha})^{\mathfrak{m}}$ and $p_{-2\alpha} \in S^2(\mathfrak{g}^{-2\alpha})^{\mathfrak{m}}$ (see §4) are the sums of the squares of the members of B_θ-orthonormal bases of $\mathfrak{g}^{-\alpha}$ and $\mathfrak{g}^{-2\alpha}$, respectively, we have

$$p_{-\alpha} = (i/3)E_{21}E_{32} \quad \text{and} \quad p_{-2\alpha} = (1/6)E_{31}^2.$$

The element $x_\alpha \in \mathfrak{a}$ (see §2) is $(1/12)(E_{11} - E_{33})$, so that $(\alpha, \alpha) =$

$B(x_\alpha, x_\alpha) = 1/12$. Hence

$$q_{-\alpha} = 24\lambda(p_{-\alpha}) = 4i(E_{21}E_{32} + E_{32}E_{21}) = 8iE_{21}E_{32} + 4iE_{31} \in \mathfrak{N}^{\mathfrak{m}}_{-\alpha}$$

and

$$q_{-2\alpha} = 6\lambda(p_{-2\alpha}) = E_{31}^2 \in \mathfrak{N}^{\mathfrak{m}}_{-\alpha},$$

in the notation of §5. We may choose $r_{-2\alpha} = E_{31} \in \mathfrak{g}^{-2\alpha}$ (see Theorem 8.6), since $\dim \mathfrak{g}^{2\alpha} = 1$ and $E_{31}^2 = q_{-2\alpha}$. By Theorem 5.1 (Case 3), $\mathfrak{N}^{\mathfrak{m}}_{-\alpha}$ is the polynomial algebra $k[q_{-\alpha}, r_{-2\alpha}]$.

Let $\nu \in \mathfrak{a}^*$. We want to determine the conical vectors in the twisted induced $\mathfrak{g}$-module $X^\nu = V^{\nu-\rho}$ induced from the subalgebra $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ of $\mathfrak{g}$, where $\rho = 2\alpha \in \mathfrak{a}^*$ and $\mathfrak{n} = \mathfrak{g}^\alpha \oplus \mathfrak{g}^{2\alpha}$ (see §2). Let x_0 be the canonical generator of X^ν. Then

$$(X^\nu)^{\mathfrak{m}} = \mathfrak{N}^{\mathfrak{m}}_{-\alpha} \cdot x_0 = k[q_{-\alpha}, r_{-2\alpha}] \cdot x_0.$$

Thus we must determine the polynomials a_0 in two variables over k such that $\mathfrak{n} \cdot (a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0) = 0$.

It is hard to guess what conical vectors should look like, but once we know, it is relatively easy to prove that they are in fact conical (in the present special case):

Lemma 9.1. *Suppose $\nu(h_\alpha) = 2l$, l a positive integer, and let*

$$x = (q_{-\alpha} - 4i(l-1)r_{-2\alpha})(q_{-\alpha} - 4i(l-3)r_{-2\alpha}) \cdots$$
$$(q_{-\alpha} + 4i(l-3)r_{-2\alpha})(q_{-\alpha} + 4i(l-1)r_{-2\alpha}) \cdot x_0$$

in X^ν. Then x is a conical vector.

Proof. Since $E_{13} = [E_{12}, E_{23}]$, $\mathfrak{g}^\alpha$ generates $\mathfrak{n}$, and so it is sufficient to show that $E_{12} \cdot x = E_{23} \cdot x = 0$. By straightforward computation, using the matrix product relation $E_{\alpha\beta}E_{\gamma\delta} = E_{\alpha\delta}$ if $\beta = \gamma$ and $= 0$ if $\beta \neq \gamma$ ($\alpha, \beta, \gamma, \delta = 1, 2, 3$), we have the following commutation relations in the universal enveloping algebra of $\mathfrak{g}$:

$$[E_{12}, q_{-\alpha}] = 4iE_{32} + 2iE_{32}h_\alpha + 4iE_{32}h', \quad [E_{12}, r_{-2\alpha}] = -E_{32},$$

$$[E_{23}, q_{-\alpha}] = 4iE_{21} + 2iE_{21}h_\alpha - 4iE_{21}h', \quad [E_{23}, r_{-2\alpha}] = E_{21}.$$

Let a be any one of the factors $q_{-\alpha} + 4ijr_{-2\alpha}$ ($j = -(l-1), -(l-3), \cdots, l-1$) appearing in the expression for x in the statement of the lemma. Then $[h_\alpha, a] = -4a$ and $[h', a] = 0$. Also $h_\alpha \cdot x_0 = (\nu - \rho)(h_\alpha)x_0 = (2l-4)x_0$ and $h' \cdot x_0 = 0$. The above commutation relations thus give

$$E_{12}\cdot x=[E_{12},(q_{-\alpha}-4i(l-1)r_{-2\alpha})](q_{-\alpha}-4i(l-3)r_{-2\alpha})$$
$$\cdots(q_{-\alpha}+4i(l-1)r_{-2\alpha})\cdot x_0$$
$$+(q_{-\alpha}-4i(l-1)r_{-2\alpha})[E_{12},(q_{-\alpha}-4i(l-3)r_{-2\alpha})]$$
$$\cdots(q_{-\alpha}+4i(l-1)r_{-2\alpha})\cdot x_0+\cdots$$
$$=(4iE_{32}+2iE_{32}h_\alpha+4iE_{32}h'+4i(l-1)E_{32})(q_{-\alpha}-4i(l-3)r_{-2\alpha})$$
$$\cdots(q_{-\alpha}+4i(l-1)r_{-2\alpha})\cdot x_0$$
$$+(q_{-\alpha}-4i(l-1)r_{-2\alpha})(4iE_{32}+2iE_{32}h_\alpha+4iE_{32}h'+4i(l-3)E_{32})$$
$$\cdots(q_{-\alpha}+4i(l-1)r_{-2\alpha})\cdot x_0+\cdots$$
$$=(4iE_{32}+2iE_{32}(-4(l-1)+2l-4)+4i(l-1)E_{32})(q_{-\alpha}-4i(l-3)r_{-2\alpha})$$
$$\cdots(q_{-\alpha}+4i(l-1)r_{-2\alpha})\cdot x_0$$
$$+(q_{-\alpha}-4i(l-1)r_{-2\alpha})(4iE_{32}+2iE_{32}(-4(l-2)+2l-4)+4i(l-3)E_{32})$$
$$\cdots(q_{-\alpha}+4i(l-1)r_{-2\alpha})\cdot x_0+\cdots$$
$$=0+0+\cdots=0.$$

A similar computation shows that $E_{23}\cdot x_0=0$. However, x must be written in the "opposite order," as

$$(q_{-\alpha}+4i(l-1)r_{-2\alpha})(q_{-\alpha}+4i(l-3)r_{-2\alpha})$$
$$\cdots(q_{-\alpha}-4i(l-3)r_{-2\alpha})(q_{-\alpha}-4i(l-1)r_{-2\alpha})\cdot x_0,$$

to make the computation exactly parallel to the above one. Q.E.D.

Remark. Because of the flexibility allowed in writing the expression for x in either order in the above proof, we could prove easily that x is conical without appealing to the difficult commutation relations in $\mathfrak{N}^-$. This flexibility is lost for Lie algebras $\mathfrak{g}$ in which the double root space $\mathfrak{g}^{2\alpha}$ is more than one-dimensional, since the "square root" $r_{-2\alpha}$ of $q_{-2\alpha}$ does not exist.

Now we turn to the uniqueness of the conical vectors.

Lemma 9.2. *Let $a_0(y,z)$ be a polynomial in two variables over k. Then $a_0(q_{-\alpha},r_{-2\alpha})\cdot x_0$ is a conical restricted weight vector in X^ν if and only if either a_0 is a nonzero scalar or else $\nu(h_\alpha)=2l$, where l is a positive integer, and a_0 is a nonzero multiple of*

$$a_l = (y - 4i(l-1)z)(y - 4i(l-3)z) \cdots (y + 4i(l-3)z)(y + 4i(l-1)z).$$

If l is even, then

$$a_l = \prod_{j=1;\, j \text{ odd}}^{l-1} (y^2 + 16j^2z^2),$$

and if l is odd,

$$a_l = y \prod_{j=2;\, j \text{ even}}^{l-1} (y^2 + 16j^2z^2).$$

Proof. Let $\mathfrak{h} = \mathfrak{g}^0 = \mathfrak{m} \oplus \mathfrak{a}$. Then $\mathfrak{h}$ is a Cartan subalgebra of $\mathfrak{g}$, and the elements y of $\mathfrak{h}$ can be written $y = y_1E_{11} + y_2E_{22} + y_3E_{33}$, where $y_i \in k$ and $y_1 + y_2 + y_3 = 0$. Define $\lambda_1, \lambda_2, \lambda_3 \in \mathfrak{h}^*$ by the formulas

$$\lambda_1(y) = y_1 - y_2, \quad \lambda_2(y) = y_2 - y_3 \quad \text{and} \quad \lambda_3(y) = y_1 - y_3.$$

Then the set R of roots of $\mathfrak{g}$ with respect to $\mathfrak{h}$ is $\{\pm\lambda_1, \pm\lambda_2, \pm\lambda_3\}$. Denoting the root spaces for $\mathfrak{g}$ with respect to $\mathfrak{h}$ by $\mathfrak{g}^{\pm\lambda_i}$, we have $\mathfrak{g}^{\lambda_1} = kE_{12}$, $\mathfrak{g}^{\lambda_2} = kE_{23}$, $\mathfrak{g}^{\lambda_3} = kE_{13}$, $\mathfrak{g}^{-\lambda_1} = kE_{21}$, $\mathfrak{g}^{-\lambda_2} = kE_{32}$ and $\mathfrak{g}^{-\lambda_3} = kE_{31}$. Let $R_+ = \{\lambda_1, \lambda_2, \lambda_3\}$, so that R_+ is a positive system in R. Then the previously defined subalgebra $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ of $\mathfrak{g}$ is the same as the Borel subalgebra $\mathfrak{b} = \mathfrak{h} \oplus \coprod \mathfrak{g}^\lambda\ (\lambda \in R_+)$, and $\mathfrak{n} = \coprod \mathfrak{g}^\lambda\ (\lambda \in R_+)$. Let $\rho' \in \mathfrak{h}^*$ be the linear functional which is the previously defined ρ on $\mathfrak{a}$ and 0 on $\mathfrak{m}$. Then $\rho' = \frac{1}{2}(\lambda_1 + \lambda_2 + \lambda_3)$, i.e., ρ' is half the sum of the positive roots of $\mathfrak{g}$ with respect to $\mathfrak{h}$. Also, define $\nu' \in \mathfrak{h}^*$ by $\nu' = \nu$ on $\mathfrak{a}$ and $\nu' = 0$ on $\mathfrak{m}$. Then the previously defined twisted induced $\mathfrak{g}$-module X^ν is the same as the Verma module associated with ν', in the sense of [2, §7.1.4]. That is, X^ν is the $\mathfrak{g}$-module induced by the character of $\mathfrak{b}$ which is $\nu' - \rho'$ on $\mathfrak{h}$ and 0 on $\mathfrak{n}$.

In order to describe the Weyl group W_R of $\mathfrak{g}$ with respect to $\mathfrak{h}$, let $\mathfrak{h}_1$ be the space of all (not necessarily traceless) 3×3 diagonal matrices and let $\mu_1, \mu_2, \mu_3 \in \mathfrak{h}_1^*$ be the basis of $\mathfrak{h}_1^*$ dual to the basis E_{11}, E_{22}, E_{33} of $\mathfrak{h}_1$. Now $\mathfrak{h}^*$ may be identified with the space of k-linear combinations of μ_1, μ_2 and μ_3, modulo the subspace $k(\mu_1 + \mu_2 + \mu_3)$. Then W_R is the group of automorphisms of $\mathfrak{h}^*$ induced by the six permutations of μ_1, μ_2 and μ_3.

Let $\nu_1 \in \mathfrak{a}^*$, and define $\nu_1' \in \mathfrak{h}^*$ to be ν_1 on $\mathfrak{a}$ and 0 on $\mathfrak{m}$. Then $x_1 \in X^\nu$ is a conical vector with restricted weight ν_1 if and only if x_1 is a (nonzero) $\mathfrak{n}$-invariant vector with weight ν_1' for the action of $\mathfrak{h}$ on X^ν. But there exists a nonzero $\mathfrak{n}$-invariant vector in X^ν with weight $\nu_2 \in \mathfrak{h}^*$ only if there exists $w \in W_R$ such that $\nu_2 + \rho' = w\nu'$ and $\nu' - (\nu_2 + \rho')$ is a nonnegative integral linear combination of the elements of R_+, by [2, Proposition

7.6.2]. Moreover, the $\mathfrak{n}$-invariant vectors in X^ν with weight ν_2 form at most a one-dimensional space, by a theorem of Verma [2, Théorème 7.6.6]. Let Z be the intersection of the conical space of X^ν with the restricted weight space corresponding to ν_1. It follows that if $Z \neq 0$, then $\dim Z = 1$, and in this case, either $\nu_1 = \nu - \rho$, or else $\nu_1 = -\nu - \rho$ and $\nu = l\alpha$ (i.e., $\nu(h_\alpha) = 2l$), where l is a nonnegative integer. Now apply Lemma 9.1. (If $l = 0$, then $\nu = 0$, $\nu_1 = -\rho$ and Z is the span of x_0.) Q.E.D.

10. Conclusions. We are now ready to combine the results of §§5, 6, 8 and 9 to remove the hypothesis "$2\alpha \notin \Sigma$" from Theorems 6.17 and 6.18.

Let $(\mathfrak{g}, \theta)$ be a semisimple symmetric Lie algebra over the field k of characteristic zero, $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ the symmetric decomposition of $(\mathfrak{g}, \theta)$, $\mathfrak{a}$ a splitting Cartan subspace of $\mathfrak{p}$, $\Sigma \subset \mathfrak{a}^*$ the corresponding restricted root system, $\Sigma_+ \subset \Sigma$ a positive system, and $\rho \in \mathfrak{a}^*$ as defined in §2. For every $\phi \in \Sigma$, define $h'_\phi \in \mathfrak{a}$ to be h_ϕ if $\dim \mathfrak{g}^\phi > 1$ (see §2) and $2h_\phi$ if $\dim \mathfrak{g}^\phi = 1$. Let s_ϕ be the Weyl reflection with respect to ϕ (see §2). Also, let q_ϕ and $q_{2\phi}$ be the elements of the universal enveloping algebra of $\mathfrak{g}$ defined in §5; if $2\phi \notin \Sigma$, take $q_{2\phi} = 0$.

Here are our main results, which generalize Theorems 6.17 and 6.18:

Theorem 10.1. *Let $\alpha \in \Sigma_+$ be a simple restricted root and $\nu \in \mathfrak{a}^*$. Let Y be the subspace of the twisted induced $\mathfrak{g}$-module X^ν spanned by the conical restricted weight vectors with restricted weights of the form $\nu - \rho + c\alpha$ $(c \in k)$; if* $\dim \mathfrak{a} = 1$, *then Y is the conical space of X^ν. Then* $\dim Y$ *is either 1 or 2. If $\nu(h'_\alpha)$ is not a positive even integer, then Y is the span of x_0, the canonical generator of X^ν. Suppose $\nu(h'_\alpha) = 2l$, l a positive integer. Then* $\dim Y = 2$. *Define the element ζ_l in the universal enveloping algebra of $\mathfrak{g}$ as follows: If* $\dim \mathfrak{g}^\alpha > 1$ *and l is even,*

$$\zeta_l = \prod_{j=1;\, j \text{ odd}}^{l-1} (q_{-\alpha}^2 + 16 j^2 q_{-2\alpha});$$

if $\dim \mathfrak{g}^\alpha > 1$ *and l is odd,*

$$\zeta_l = q_{-\alpha} \prod_{j=2;\, j \text{ even}}^{l-1} (q_{-\alpha}^2 + 16 j^2 q_{-2\alpha});$$

and if $\dim \mathfrak{g}^\alpha = 1$, *$\zeta_l = f^l$, where f is a nonzero element of $\mathfrak{g}^{-\alpha}$. Then Y has basis $\{x_0, \zeta_l \cdot x_0\}$, and $\zeta_l \cdot x_0$ is a conical restricted weight vector in X^ν with restricted weight $s_\alpha \nu - \rho$.*

Theorem 10.2. *Let α be a simple restricted root, let $\mu, \nu \in \mathfrak{a}^*$, and suppose that $\mu - \nu$ is of the form $c\alpha$ $(c \in k)$. (If* $\dim \mathfrak{a} = 1$, *then this is automa-*

tic.) Then $\mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu)$ *is at most one-dimensional, and* $\dim \mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu) = 1$ *if and only if either* $\mu = \nu$, *or else* $\mu = s_\alpha \nu$ *and* $\nu(h'_\alpha)$ *is a nonnegative even integer. Moreover,* $\dim \mathrm{Hom}_{\mathfrak{g}}(X^\mu, X^\nu) = 1$ *if and only if* X^μ *is isomorphic to a* $\mathfrak{g}$*-submodule of* X^ν.

Proof. Theorem 10.2 follows from Theorem 10.1, just as in the proof of Theorem 6.18. To prove Theorem 10.1, note first that the case $2\alpha \notin \Sigma$ is covered in Theorem 6.17. Suppose that $2\alpha \in \Sigma$. It is clearly sufficient to assume now that k is algebraically closed. By Lemma 6.16, $Y = (\mathfrak{N}^{\mathfrak{m}}_{-\alpha} \cdot x_0)^{\mathfrak{n}}$. Moreover, $\mathfrak{N}^{\mathfrak{m}}_{-\alpha}$ is the polynomial algebra $k[q_{-\alpha}, q_{-2\alpha}]$ if $\dim \mathfrak{g}^{2\alpha} > 1$ and $\mathfrak{N}^{\mathfrak{m}}_{-\alpha}$ is the polynomial algebra $k[q_{-\alpha}, r_{-2\alpha}]$ if $\dim \mathfrak{g}^{2\alpha} = 1$, by Theorem 5.1; here $r_{-2\alpha} \in \mathfrak{g}^{-2\alpha}$ and $r^2_{-2\alpha} = q_{-2\alpha}$ (such an element exists since k is algebraically closed). Hence Y is the set of $\mathfrak{m} \oplus \mathfrak{n}$-invariants in X^ν of the form $a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ if $\dim \mathfrak{g}^{2\alpha} = 1$ and of the form $a_0(q_{-\alpha}, q_{-2\alpha}) \cdot x_0$ if $\dim \mathfrak{g}^{2\alpha} > 1$, where a_0 ranges through the polynomials in two variables over k. The stage is set for the application of the transfer principle for conical vectors (Theorem 8.6). Suppose that $\dim \mathfrak{g}^{2\alpha} = 1$, and that $a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ is a conical vector. If $\nu(h_\alpha)$ is not a positive even integer, then a_0 is a nonzero scalar, by the last part of Theorem 8.6, combined with Lemma 9.2. Suppose now that $\nu(h_\alpha) = 2l$, where l is a positive integer. Then the same two results show that $a_0(q_{-\alpha}, r_{-2\alpha}) \cdot x_0$ is a (nonzero) linear combination of x_0 and $\zeta_l \cdot x_0$, in the notation of the theorem. Conversely, $\zeta_l \cdot x_0$ is, in fact, a conical vector, again by Theorem 8.6 and Lemma 9.2 (or Lemma 9.1). This proves the present theorem in case $\dim \mathfrak{g}^{2\alpha} = 1$. If $\dim \mathfrak{g}^{2\alpha} > 1$, the theorem follows from the same argument, this time using the first part of Theorem 8.6. Note that since the polynomials a_l in Lemma 9.2 are polynomials in y and z^2, the space Y has the same description whether $\dim \mathfrak{g}^{2\alpha} = 1$ or $\dim \mathfrak{g}^{2\alpha} > 1$. Q.E.D.

Remark. (Cf. the Remark following Theorem 6.17.) In the notation of Theorem 10.1, $\nu(h'_\alpha)$ is a nonnegative even integer if and only if X^ν contains an $\mathfrak{m}$-invariant restricted weight vector with restricted weight $s_\alpha\nu - \rho$, or equivalently, a conical restricted weight vector with restricted weight $s_\alpha \nu - \rho$. But in general not every $\mathfrak{m}$-invariant restricted weight vector with restricted weight $s_\alpha \nu - \rho$ is conical.

Remark. If $\dim \mathfrak{a} = 1$ and $\dim \mathfrak{g}^\alpha > 1$, then $\nu(h'_\alpha) = \nu(h_\alpha)$ is a nonnegative even integer if and only if ν is a nonnegative integral multiple of the unique simple restricted root α.

BIBLIOGRAPHY

1. I. N. Bernšteĭn, I. M. Gel'fand and S. I. Gel'fand, (a) *Structure of representa-*

tions generated by highest weight vectors, Funkcional. Anal. i Priložen. 5 (1971), 1–9 = Functional Anal. Appl. 5 (1971), 1–8.

(b) *Differential operators on the fundamental affine space*, Dokl. Akad. Nauk SSSR 195 (1970), 1255–1258. (Russian) MR 43 #3402.

2. J. Dixmier, *Algèbres enveloppantes*, Gauthier-Villars, Paris, 1974.

3. M. Duflo, *Représentations irréductibles des groupes semi-simples complexes* (to appear).

4. Harish-Chandra, *Representations of semisimple Lie groups.* II, Trans. Amer. Math. Soc. 76 (1954), 26–65. MR 15, 398.

5. S. Helgason, (a) *A duality for symmetric spaces with applications to group representations*, Advances in Math. 5 (1970), 1–154. MR 41 #8587.

(b) *Analysis on Lie groups and homogeneous spaces*, CBMS Regional Conference Series in Math., no. 14, Amer. Math. Soc., Providence, R. I., 1972. MR 47 #5179.

6. B. Kostant, *On the existence and irreducibility of certain series of representations*, Publ. 1971 Summer School in Math., edited by I. M. Gel'fand, Bolyai-Janós Math. Soc., Budapest (to appear).

7. J. Lepowsky, (a) *Algebraic results on representations of semisimple Lie groups*, Trans. Amer. Math. Soc. 176 (1973), 1–44.

(b) *On the Harish-Chandra homomorphism*, Trans. Amer. Math. Soc. 208 (1975), 193–218.

(c) *Uniqueness of embeddings of certain induced modules* (to appear).

(d) *On the uniqueness of conical vectors* (to appear).

(e) *A generalization of H. Weyl's "unitary trick"*, Trans. Amer. Math. Soc. (to appear).

8. G. D. Mostow, *Rigidity of locally symmetric spaces*, Ann. of Math. Studies, no. 78, Princeton Univ. Press, Princeton, N. J., 1973.

9. C. Rader, *Spherical functions on semisimple Lie groups*, Thesis and unpublished supplements, University of Washington, 1971.

10. D.-N. Verma, (a) *Structure of certain induced representations of complex semisimple Lie algebras*, Thesis, Yale University, 1966.

(b) *Structure of certain induced representations of complex semisimple Lie algebras*, Bull. Amer. Math. Soc. 74 (1968), 160–166; errata, p. 628. MR 36 #1503; #5182.

11. N. R. Wallach, *Harmonic analysis on homogeneous spaces*, Pure and Appl. Math., vol. 19, Dekker, New York, 1973.

12. M. Hu, *Determination of the conical distributions for rank one symmetric spaces*, Thesis, Massachusetts Institute of Technology, 1973.

DEPARTMENT OF MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, CONNECTICUT 06520

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

THE GENERALIZED MARTIN'S MINIMUM PROBLEM AND ITS APPLICATIONS IN SEVERAL COMPLEX VARIABLES

BY

SHOZO MATSUURA

ABSTRACT. The objectives of this paper are to generalize the Martin's $\mathfrak{L}^2$-minimum problem under more general additional conditions given by bounded linear functionals in a bounded domain D in C^n and to apply this problem to various directions.

We firstly define the new ith biholomorphically invariant Kähler metric and the ith representative domain ($i = 0, 1, 2, \ldots$), and secondly give estimates on curvatures with respect to the Bergman metric and investigate the asymptotic behaviors via an A-approach on the curvatures about a boundary point having a sort of pseudoconvexity.

Further, we study (i) the extensions of some results recently obtained by K. Kikuchi on the Ricci scalar curvature, (ii) a minimum property on the reproducing subspace-kernel in $\mathfrak{L}^2_{(m)}(D)$, and (iii) an extension of the fundamental theorem of K. H. Look.

1. Introduction. The Bergman's minimum problem [3] with respect to $\mathfrak{L}^2(D)$ under some additional conditions has been extended by W. T. Martin [15] as the following (originally posed by W. Wirtinger [21]): Find the function $f(z)$ (belonging to $\mathfrak{L}^2(D)$ or $\mathfrak{L}^2_{X,t}(D)$) which minimizes the Lebesgue square integral $(Q - f, Q - f)_D$ for a given function $Q(z, \bar{z}) \in L^2(D)$. Here $L^2(D)$ and $\mathfrak{L}^2(D)$ denote the classes of square integrable and of square integrable holomorphic functions in a bounded domain D, respectively. $\mathfrak{L}^2_{X,t}(D)$ denotes the class $\{f(z) \in \mathfrak{L}^2(D) | f(t) = X,\ t \in D\}$.

In §3, under more general additional conditions using bounded linear functionals we shall get the generalized Martin's theorem, which includes the cases of Bergman [3], Martin [15] and others [17], [19], [20].

As an application of the minimum problem, in §4 we shall define the

Received by the editors March 19, 1974.

AMS (MOS) subject classifications (1970). Primary 32H05, 32H10, 32H15; Secondary 53A55.

Key words and phrases. Bergman kernel function, Bergman metric, biholomorphically invariant, representative domain, bounded linear functional, holomorphic bisectional curvature, Ricci tensor, strictly pseudoconvex, classical Cartan domains.

interesting quantities $\Omega_D^{(i)}(z)$ and $\widetilde{\Omega}_D^{(i)}(z)$ $(i = 0, 1, 2, \ldots)$ which have a sort of positivity and play important roles throughout this paper. Using these, we shall define the new ith biholomorphically invariant Kähler metric $(ds_D^{(i)})^2 = \partial_z^* \partial_z \log \det \widetilde{K}_D^{(i)}(z, \bar{z})$ and the ith representative domain $(i = 0, 1, 2, \ldots)$, where the biholomorphically relative invariants $\widetilde{K}_D^{(i)}(z, \bar{z})$ $(i = 0, 1, 2, \ldots)$ are constructed by the Bergman kernel function of a bounded domain D and its derivatives. In particular, $(ds_D^{(0)})^2$ and $(ds_D^{(1)})^2$ coincide with the Bergman metric [3] and the Fuks metric [8], respectively, and the 0th representative domain coincides with the Bergman representative domain.

In §§5, 6 and 7, using the results of §§3 and 4, we shall give various estimations (Theorems 5.1 and 5.2) on the holomorphic "bisectional" curvature $R_D(z; u, v)$, the Ricci curvature $C_D(z; u)$ and the Ricci scalar curvature $S_D(z)$ of a bounded domain D with respect to the Bergman metric and generalize the results obtained by S. Bergman [1], [2], [3], B. A. Fuks [6], [7], [8] and others. For our purpose, the quantity $\Omega_D^{(2)}(z)$ and "the method of minimum integral" [3], [7] are used effectively.

In the case of C^2, the asymptotic behaviors of the Bergman kernel function $k_D(z, \bar{z})$ and related biholomorphic invariants about a boundary point Q of a domain D such that the Levi determinant $L(\phi)$ is positive at Q have been studied minutely by S. Bergman [1] and B. A. Fuks [6], [7], [8]. But in the case of C^n $(n \geq 3)$, few results are known (see Chalmers [4], Hörmander [9]). On the asymptotic behaviors of the curvatures of a bounded domain D in C^n about a boundary point Q at which D is strictly pseudoconvex globally representable [4] and has the normal analytic hypersurface h (through Q) lying entirely outside itself, in §7 we shall prove that, using a sort of domains of comparison due to B. Chalmers [4], $R_D(z; u)$ $(\equiv R_D(z; u, v))$, $C_D(z; u)$ and $S_D(z)$ tend to $-2/(n+1)$, -1 and $-n$ via an A-approach: $z \longrightarrow Q$, respectively.

In §8, some results recently obtained by K. Kikuchi [12] with respect to the Ricci scalar curvature as an application of the theorem of E. Hopf are extended.

In §9, using the minimum problem with the condition that $Q(z, \bar{z}) \equiv Q(z) = k_D(z, \bar{t}) \in \mathfrak{L}^2(D)$, where $k_D(z, \bar{t})$ denotes the Bergman kernel function of D, we shall show that the reproducing kernel function of a subspace $\mathfrak{L}^2_{(m)}(D)$ of $\mathfrak{L}^2(D)$ (see [5], [18]) has a sort of minimum property and give another expression of this kernel given in [5].

Finally, in §10 a neat proof and an extension of the fundamental theorem (I) of K. H. Look [14] are given.

2. Preliminaries. Throughout this paper we shall use, as far as possible, matrix representations, which give us available perspectives. For a matrix A, $\bar{A}$, A^T and A^* denote the conjugate, the transposed and the conjugate transposed matrices of A, respectively. The symbol $\times$ shows the Kronecker product and $[A]^k$ denotes $A \times \cdots \times A$ (k-times).

Let D be a bounded schlicht domain in C^n and $z \equiv (z_1, \ldots, z_n)^T$ be a complex $n \times 1$ vector variable in D. For the differential operator $D_z \equiv \partial/\partial z \equiv (\partial/\partial z_1, \ldots, \partial/\partial z_n)$ ($D_z^* \equiv \partial/\partial z^* \equiv (\partial/\partial \bar{z})^T$), we shall define two sorts of the kth order differential operators with respect to z as follows:

$$[D_z]^k \equiv [\partial/\partial z]^k \equiv (\partial/\partial z) \times \cdots \times (\partial/\partial z) \qquad (1 \times n^k \text{ vector})$$

and its contraction

$$\begin{aligned} D_z^k &\equiv \partial^k/\partial z^k \\ &\equiv (\partial^k/\partial z_1^k, \ldots, (k!/k_1! \cdots k_n!)\partial^k/\partial z_1^{k_1} \cdots \partial z_n^{k_n}, \ldots, \partial^k/\partial z_n^k) \end{aligned}$$

$(1 \times {}_nH_k$ vector), where $\sum_{j=1}^n k_j = k$ and the arrangement of $\{k_1, \ldots, k_n\}$ is lexicographical. Using these operators, the kth order derivatives of a matrix function $F(z, \bar{z}) \equiv (f_{pq}(z, \bar{z}))$ with respect to z are defined by

$$[D_z]^k F(z, \bar{z}) \equiv [D_z]^k \times F(z, \bar{z}) \equiv ([D_z]^k \times f_{pq}(z, \bar{z}))$$

and

$$D_z^k F(z, \bar{z}) \equiv D_z^k \times F(z, \bar{z}) \equiv (D_z^k \times f_{pq}(z, \bar{z})).$$

If we define the contracted kth power of an $n \times 1$ vector $u \equiv (u_1, \ldots, u_n)^T$ as

$$u^k \equiv (u_1^k, \ldots, u_1^{k_1} \cdots u_n^{k_n}, \ldots, u_n^k)^T,$$

it holds that, for a scalar function $f(z, \bar{z})$,

$$(D_z^k f(z, \bar{z}))u^k = ([D_z]^k f(z, \bar{z}))[u]^k.$$

The total differential of a matrix function $F(z, \bar{z})$ ($r \times s$ type) is defined by

$$dF(z, \bar{z}) \equiv \partial_z F + \partial_z^* F \equiv (D_z F)(dz \times E_s) + (dz^* \times E_r)(D_z^* F),$$

where $dz \equiv (dz_1, \ldots, dz_n)^T$ and E_k denotes the $k \times k$ unit matrix.

In the following, we shall use some available formulas with respect to matrices, derivatives and differentials without proof [12], [16], [17]:

$$D_z(AB) = (D_zA)(E_n \times B) + A(D_zB) \tag{2.1}$$

(A, B are $k \times l$, $l \times m$ matrices, respectively),

$$D_z(A \times B) = (D_zA) \times B + (A \times D_zB)(\tilde{E}_{ln} \times E_q) \tag{2.2}$$

(A, B are $k \times l$, $p \times q$ matrices respectively and

$$\tilde{E}_{ln} = \begin{pmatrix} e_{11} & \cdots & e_{l1} \\ \vdots & & \vdots \\ e_{1n} & \cdots & e_{ln} \end{pmatrix},$$

where e_{ij} ($i = 1, \ldots, l$; $j = 1, \ldots, n$) is an $l \times n$ matrix which has 1 as (i, j)-element and 0's elsewhere),

$$\begin{aligned} \partial_z(A^{-1}) &= -A^{-1}(\partial_zA)A^{-1} = -A^{-1}(D_zA)(dz \times A^{-1}), \\ D_z(A^{-1}) &= -A^{-1}(D_zA)(E_n \times A^{-1}) \end{aligned} \tag{2.3}$$

(A is a $k \times k$ regular matrix) and

$$\partial_z \log \det A = \mathrm{Sp}(A^{-1}\partial_zA) = \mathrm{Sp}\{A^{-1}(D_zA)(dz \times E_k)\} \tag{2.4}$$

(A is a $k \times k$ regular matrix and Sp denotes the trace symbol). By (2.3) and (2.4) we have the following lemma.

Lemma 2.1. *For a $k \times k$ regular matrix function $A(z, \bar{z})$ we have*

$$\partial_z^*\partial_z \log\det A = \mathrm{Sp}\{(dz^* \times E_k)(A_{11} - A_{10}A^{-1}A_{01})(dz \times E_k)A^{-1}\}, \tag{2.5}$$

where A_{11} denotes $D_z^*D_zA$, etc.

Let $H(D)$ be the class of holomorphic matrix functions of all types in D and $BH(D)$ be the subclass of $H(D)$ defined by

$$BH(D) \equiv \{f(z) \equiv (f_1(z), \ldots, f_n(z))^T \in H(D) | J_f(z) \neq 0 \text{ in } D \subset C^n\},$$

where $J_f(z)$ denotes the Jacobian determinant $\det(df(z)/dz)$ ($\equiv \det(D_zf(z))$). We call each element belonging to $BH(D)$ a biholomorphic mapping, which is locally one-to-one in D. The subclass $\mathfrak{L}^2(D)$ of $H(D)$, which denotes the class of square Lebesgue integrable holomorphic functions in a bounded

domain D, makes a complete Hilbert space with the Bergman reproducing kernel function $k_D(z, \bar{t})$.

3. General minimum problem. In this section, we shall generalize the results of S. Bergman [3], W. Wirtinger [21], W. T. Martin [15] and others for a given complex-valued $r \times 1$ vector function $Q(z, \bar{z}) \in L^2(D)$ and a general class (with more general additional conditions)

$$\mathfrak{L}_K^2(D) \equiv \{f(z)(r \times 1 \text{ type}) \in \mathfrak{L}^2(D) | \mathfrak{L}f = K, \mathfrak{L} \in BL(D)\},$$

where $BL(D)$ denotes the class of all types of bounded linear functional matrices (see [5]) and K denotes a given constant matrix of the same type as $\mathfrak{L}f$.

Theorem 3.1. *For a given $r \times 1$ vector function $Q(z, \bar{z}) \in L^2(D)$ in a bounded domain D, the minimizing function $M_{D,Q}^K(z) \in \mathfrak{L}_K^2(D)$, which minimizes the Lebesgue square integral*

$$I(Q, f) \equiv (Q - f, Q - f)_D \equiv \operatorname{Sp} \int_D (Q(\zeta, \bar{\zeta}) - f(\zeta))(Q(\zeta, \bar{\zeta}) - f(\zeta))^* \omega_\zeta \tag{3.1}$$

under an additional condition

$$\mathfrak{L}f = K \ (K\text{: a given constant matrix, } \mathfrak{L} \in BL(D)) \tag{3.2}$$

with the condition $\det(\Phi^*\Phi) \neq 0$ *for* $\Phi \equiv \mathfrak{L}\phi_D$ $(\phi_D(z) \equiv (\phi_1(z), \phi_2(z), \ldots)^T$: *an orthonormal system in* $\mathfrak{L}^2(D))$, *is given by*

$$M_{D,Q}^K(z) = \{B + (K - B\Phi)(\Phi^*\Phi)^{-1}\Phi^*\}\phi_D(z) \in \mathfrak{L}_K^2(D) \tag{3.3}$$

and also the minimum value of $I(Q, f)$ is given by

$$\lambda_{D,Q}^K = \operatorname{Sp}\{BB^* - B\Phi(\Phi^*\Phi)^{-1}\Phi^*B^* + K(\Phi^*\Phi)^{-1}K^*\}, \tag{3.4}$$

where ω_ζ denotes the Euclidean volume element $\Pi_{j=1}^n d\bar{\zeta}_j \wedge d\zeta_j/(2\sqrt{-1})^n$ and

$$B \equiv (b_{ij}) = \int_D Q(\zeta, \bar{\zeta})\phi_D^*(\zeta)\omega_\zeta. \tag{3.5}$$

Proof. Given a sufficiently large real number M, we consider a class $G \equiv \{f(z) \in \mathfrak{L}_K^2(D) | \int_D |f(z)|^2 \omega_z \leq M < +\infty\}$. G becomes a compact family, and it is known that there exists a minimizing function $M_D^K(z) \equiv M_{D,Q=0}^K(z) \in$

$\mathfrak{L}_K^2(D)$ which minimizes the integral $I(0, f) = \int_D |f(z)|^2 \omega_z$, where $M_D^K(z)$ is given by $K(\Phi^*\Phi)^{-1}\Phi^*\phi_D(z)$ and $\det(\Phi^*\Phi) \neq 0$ (see [3]).

Now, we will follow the procedure of the proof essentially due to Martin [15]. Let $M_{D,Q}^K(z)$ be the minimizing function belonging to $\mathfrak{L}_K^2(D)$, then, using an orthonormal system $\phi_D(z)$ in D, we can set $M_{D,Q}^K(z) = A\phi_D(z)$, where $A \equiv (a_{ij}) = \int_D M_{D,Q}^K(\zeta)\phi_D^*(\zeta)\omega_\zeta$ denotes the Fourier coefficient $r \times \infty$ matrix to be determined. Noting that $\mathfrak{L}M_{D,Q}^K = A\mathfrak{L}\phi_D = A\Phi$, if we set

$$I(A) \equiv (Q - M_{D,Q}^K, Q - M_{D,Q}^K)_D - \mathrm{Sp}\{(A\Phi - K)\Lambda + \Gamma^*(\Phi^*A^* - K^*)\},$$

where $\Lambda = (\lambda_{ij})$ and $\Gamma = (\gamma_{ij})$ ($i = 1, \ldots, p$; $j = 1, \ldots, r$ and p denotes the number of the columns of K) are the Lagrangian multipliers, as necessary conditions we must have the Euler's conditions

$$\partial I(A)/\partial a_{ij} = \bar{a}_{ij} - \bar{b}_{ij} - (\Phi\Lambda)_{ij} = 0, \quad \text{i.e., } A^* = B^* + \Phi\Lambda$$

and

$$\partial I(A)/\partial \bar{a}_{ij} = a_{ij} - b_{ij} - (\Gamma^*\Phi^*)_{ij} = 0, \quad \text{i.e., } A = B + \Gamma^*\Phi^*,$$

where $i = 1, \ldots, r$; $j = 1, 2, \ldots$ and $(\Phi\Lambda)_{ij}$ denotes the (i, j)-element of $\Phi\Lambda$. Hence we have $\Phi\Lambda = \Phi\Gamma$. But since $\det(\Phi^*\Phi) \neq 0$ holds in a bounded domain, we obtain $\Lambda = \Gamma$. On the other hand, as we have $K = A\Phi = (B + \Lambda^*\Phi^*)\Phi = B\Phi + \Lambda^*(\Phi^*\Phi)$, we get

$$A = B + \Lambda^*\Phi^* = B + (K - B\Phi)(\Phi^*\Phi)^{-1}\Phi^*.$$

Therefore, we must have (3.3) belonging to $\mathfrak{L}_K^2(D)$.

In order to prove that $M_{D,Q}^K(z)$ is the minimizing function required, let us consider the class $\mathfrak{L}_0^2(D) \equiv \{g(z) \in \mathfrak{L}^2(D) | g(z) = C\phi_D(z), C\Phi = 0\}$. If we set $F(z) = M_{D,Q}^K(z) + g(z)$ for each $g(z) \in \mathfrak{L}_0^2(D)$, then it is easily shown that $F(z)$ is an arbitrary function belonging to $\mathfrak{L}_K^2(D)$. It follows from term-by-term integrability (see [15]) that

$$\begin{aligned}
&\int_D \{Q(\zeta, \bar{\zeta}) - M_{D,Q}^K(\zeta)\} g^*(\zeta)\omega_\zeta \\
&\quad = \int_D Q(\zeta, \bar{\zeta})\phi_D^*(\zeta)\omega_\zeta C^* - A \int_D \phi_D(\zeta)\phi_D^*(\zeta)\omega_\zeta C^* \\
&\quad = BC^* - AC^* = BC^* - (B + \Lambda^*\Phi^*)C^* = -\Lambda^*(C\Phi)^* = 0,
\end{aligned}$$

where $\int_D \phi_D(\zeta)\phi_D^*(\zeta)\omega_\zeta = E_\infty$. Hence we obtain

$$(Q - F, Q - F)_D = (Q - M^K_{D,Q}, Q - M^K_{D,Q})_D + (g, g)_D$$
$$- 2 \operatorname{Re} \operatorname{Sp} \int_D (Q - M^K_{D,Q}) g^* \omega_\zeta$$
$$= (Q - M^K_{D,Q}, Q - M^K_{D,Q})_D + (g, g)_D > (Q - M^K_{D,Q}, Q - M^K_{D,Q})_D$$

for any $g(z) \not\equiv 0$. This completes the proof.

Remark 3.1. In Theorem 3.1 it is easily verified that the minimizing function without an additional condition (3.2) is given by

$$M_{D,Q}(z) = B\phi_D(z) = \int_D Q(\zeta, \bar{\zeta}) k_D(z, \bar{\zeta}) \omega_\zeta, \qquad k_D(z, \bar{\zeta}) \equiv \phi^*_D(\zeta)\phi_D(z),$$

where $k_D(z, \bar{\zeta})$ denotes the Bergman kernel function [15].

In the case that $Q(z, \bar{z}) \equiv 0$ in D, the minimizing function $M^K_D(z) \equiv M^K_{D,Q=0}(z)$ and the minimum value $\lambda^K_D \equiv \lambda^K_{D,Q=0}$ are expressed in terms of the kernel function of D and its derivatives [3], [20].

Let $\mathfrak{L}_{(m)} \equiv (\mathfrak{L}_1, \dots, \mathfrak{L}_m)$ be an element of $BL(D)$ and $\mathfrak{L}_{(m),t}$ and $\mathfrak{L}_{k,t}$ be the bounded linear functionals $\mathfrak{L}_{(m)}$ and $\mathfrak{L}_k$ evaluated at a point $t \in D$. $\mathfrak{L}^2_{K(m)}(D)$ and $\mathfrak{L}^2_{K(m),t}(D)$ denote the subclasses of $\mathfrak{L}^2(D)$ such that $\{f(z) \in \mathfrak{L}^2(D) | \mathfrak{L}_{(m)} f = K(m) \equiv (A_1, \dots, A_m)\}$ and $\{f(z) \in \mathfrak{L}^2(D) | \mathfrak{L}_{(m),t} f = K(m)\}$, respectively. Here $\mathfrak{L}_{k,t} f$ denotes, say, any one of $f(t)$, $D^k_z f(t)$, $(D^k_z f(t))u^k$, $\int_0^t f(z)dz$ and $\int_D f(z)\omega_z$ and so on.

Theorem 3.1 gives the generalizations of (i) [15], (ii) [15, (5.5)], (iii) [3], [20] and (iv) [19] under the additional conditions

(i)′ $Q(z, \bar{z}) \in L^2(D)$, $\mathfrak{L}_{(m),t} f = K(m)$, $t \in D$,

(ii)′ $Q(z, \bar{z}) \in L^2(D)$, $\mathfrak{L}_{(m)} f \equiv (\mathfrak{L}_{1,t_1}, \dots, \mathfrak{L}_{m,t_m}) f \equiv (f(t_1), \dots, f(t_m)) = K(m)$, $t_k \in D$ $(k = 1, \dots, m)$,

(iii)′ $Q(z, \bar{z}) \equiv 0$, $\mathfrak{L}_{(m),t} f \equiv (\mathfrak{L}_{1,t}, \dots, \mathfrak{L}_{m,t}) f = K(m)$, where $\mathfrak{L}_{k,t} f \equiv (D^k_z f(t))u_k$ (u_k denotes a constant ${}_nH_k \times i_k$ matrix $(k = 1, \dots, m)$ and i_k denotes an arbitrary integer belonging to $\{1, 2, \dots, {}_nH_k\}$ (${}_nH_k$: repeated combination), and

(iv)′ $Q(z, \bar{z}) \equiv 0$, $\mathfrak{L}_{(2),t} f \equiv (\mathfrak{L}_{1,t}, \mathfrak{L}_{2,t}) f \equiv (f(t), \int_D f(\zeta)\omega_\zeta) = K(2)$, respectively. □

In the following we shall use the abbreviated notations $f_{ij}(z, \bar{x})$ and $f_{[ij]}(z, \bar{x})$ instead of $(D^*_z)^i (D_z)^j f(z, \bar{x})$ and $[D^*_z]^i [D_z]^j f(z, \bar{x})$, respectively. In particular, $f_{00}(z, \bar{x})$ denotes $f(z, \bar{x})$ and $f_{ij}(a, \bar{b})$ means $f_{ij}(z, \bar{x})|_{z=a, x=b}$.

In a bounded domain D, the Bergman kernel function $k_D(z, \bar{z})$ is positive and relatively invariant under $BH(D)$ and $\log k_D(z, \bar{z})$ defines a strongly

plurisubharmonic function. Therefore, an absolutely invariant Kähler metric under $BH(D)$, which is called the Bergman metric, is defined as

$$ds_D^2 \equiv dz^* T_D(z, \bar{z}) dz, \tag{3.6}$$

where the fundamental tensor

$$\begin{aligned} T_D(z, \bar{t}) &\equiv D_t^* D_z \log k_D(z, \bar{t}) \\ &= \{k(z, \bar{t}) \times k_{11}(z, \bar{t}) - k_{10}(z, \bar{t}) \times k_{01}(z, \bar{t})\}/k^2(z, \bar{t}) \end{aligned} \tag{3.7}$$

belongs to $H(D \times D^*)$ when $k(z, \bar{t}) \equiv k_D(z, \bar{t}) \neq 0$ and has the relative invariancy under $BH(D)$, where $k_{ij}(z, \bar{t})$ denotes $k_{D,ij}(z, \bar{t})$, etc.

The following lemma is known [2], [3], [7].

Lemma 3.1. *We consider the case that* $Q(z, \bar{z}) \equiv 0$ *in* D.

(i) *Under* $\mathfrak{L}_{(2),t} f \equiv (f(t), D_z f(t)) = K(2) \equiv (A_1, A_2)$ *we have*

$$M_D^{K(2)}(z, t) = (A_1, A_2) \begin{pmatrix} k & k_{01} \\ k_{10} & k_{11} \end{pmatrix}^{-1} \begin{pmatrix} k(z, \bar{t}) \\ k_{10}(z, \bar{t}) \end{pmatrix}, \tag{3.8}$$

where

$$\begin{pmatrix} k & k_{01} \\ k_{10} & k_{11} \end{pmatrix}^{-1} = \begin{pmatrix} 1/k + k_{01}(kT)^{-1}k_{10}/k^2, & -k_{01}(kT)^{-1}/k \\ -(kT)^{-1}k_{10}/k, & (kT)^{-1} \end{pmatrix}, \tag{3.9}$$

$k_{ij} \equiv k_{D,ij}(t, \bar{t})$ *and* $T \equiv T_D(t, \bar{t})$.

In particular, under $K(2) \equiv (0, E_n)$ *we have*

$$\lambda_D^{0E_n}(t) = \mathrm{Sp}\,(kT)^{-1}. \tag{3.10}$$

(ii) *Under* $\mathfrak{L}_{(2),t} f \equiv (f(t), D_z f(t)u) = K(2) \equiv (0, 1)$ *we have*

$$\lambda_D^{01}(t) \equiv \lambda_D^{(2)}(u) = 1/ku^* Tu. \tag{3.11}$$

(iii) *Under* $\mathfrak{L}_{(1),t} f \equiv (f(t)) = K(1) \equiv (1)$ *we have*

$$\lambda_D^1(t) \equiv \lambda_D^{(1)} = 1/k. \tag{3.12}$$

(iv) *Under* $\mathfrak{L}_{(3),t} f \equiv (f(t), D_z f(t), [D_z]^2 f(t)(u \times v)) = K(3) \equiv (0, \ldots, 0, 1)$ *we have*

$$\lambda_D^{001}(t) \equiv \lambda_D^{(3)}(u, v) = \lambda_D^{(2)}(u)\lambda_D^{(2)}(v)/\lambda_D^{(1)}(u \times v)^*\Omega_D^{(2)}(t)(u \times v), \tag{3.13}$$

where $\Omega_D^{(2)}(t)$ *is defined in* (4.5).

Remark 3.2. For a regular matrix $A \equiv \begin{pmatrix} K & L \\ M & N \end{pmatrix}$, if K and $Z \equiv N - MK^{-1}L$ are regular, then we have

$$A^{-1} = \begin{pmatrix} K^{-1} + XZ^{-1}Y, & -XZ^{-1} \\ -Z^{-1}Y, & Z^{-1} \end{pmatrix}, \tag{3.14}$$

where $X \equiv K^{-1}L$ and $Y \equiv MK^{-1}$.

4. New invariant Kähler metrics.

Definition 4.1. We define the two sorts of matrices:

$$\widetilde{K}_D^{(i)}(z, \bar{z}) \equiv \begin{pmatrix} k & k_{01} \cdots & k_{0i} \\ k_{10} & & \vdots \\ \vdots & & \vdots \\ k_{i0} & \cdots & k_{ii} \end{pmatrix} \equiv \begin{pmatrix} \widetilde{K}_D^{(i-1)}(z, \bar{z}), & \widetilde{P}^{(i-1)} \\ (\widetilde{P}^{(i-1)})^*, & k_{ii} \end{pmatrix}, \tag{4.1}$$

$i = 0, 1, 2, \ldots$, and

$$K_D^{(i)}(z, \bar{z}) \equiv \begin{pmatrix} \widetilde{K}_D^{(i-1)}(z, \bar{z}), & P^{(i-1)} \\ (P^{(i-1)})^*, & (k_{i-1,i-1})_{11} \end{pmatrix}, \quad P^{(i-1)} \equiv \begin{pmatrix} (k_{0,i-1})_{01} \\ \vdots \\ (k_{i-1,i-1})_{01} \end{pmatrix}, \tag{4.2}$$

where $(k_{pq})_{01}$ denotes $D_z\{(D_z^*)^p D_z^q k_D(z, \bar{z})\}$, etc., and $\widetilde{K}_D^{(i)}$ and $K_D^{(i)}$ are $s(i) \times s(i)$ and $\{s(i-1) + nt(i-1)\} \times \{s(i-1) + nt(i-1)\}$ matrices, respectively. Here $t(i)$ and $s(i)$ denote ${}_nH_i$ and $\sum_{k=0}^{i} t(k)$ $(= \binom{n+i}{i})$, respectively.

Lemma 4.1. *In a bounded domain D, we have*

$$\det \widetilde{K}_D^{(i)}(z, \bar{z}) > 0, \qquad \det K_D^{(i)}(z, \bar{z}) \equiv 0 \qquad (i \geq 2 \textit{ in the latter}), \tag{4.3}$$

$$\widetilde{\Omega}_D^{(i)}(z) \equiv \{k_{ii} - (\widetilde{P}^{(i-1)})^*(\widetilde{K}_D^{(i-1)})^{-1}\widetilde{P}^{(i-1)}\}/k > 0 \tag{4.4}$$

and

$$\begin{aligned} &\Omega_D^{(i)}(z) \equiv \{(k_{i-1,i-1})_{11} - (P^{(i-1)})^*(\widetilde{K}_D^{(i-1)})^{-1}P^{(i-1)}\}/k, \\ &(u^* \times E_{t(i-1)})\Omega_D^{(i)}(z)(u \times E_{t(i-1)}) > 0 \end{aligned} \tag{4.5}$$

for $i \geq 0$, *where* $t(i-1) = {}_nH_{i-1}$.

Proof. $\det \widetilde{K}_D^{(0)}(z, \bar{z}) = k_D(z, \bar{z}) > 0$ in D is clear. Since k_D^{-1} exists, then we have $\det K_D^{(1)} \equiv k_D^{n+1}(z, \bar{z}) \det T_D(z, \bar{z}) > 0$ in D.

Now, let us suppose that $\det \widetilde{K}_D^{(i-1)}(z, \bar{z}) > 0$ in D. Under the condition $(f(t), D_z f(t), \ldots, (D_z^i f(t))v) = (0, \ldots, 0, 1) \equiv K(i+1)$, where v denotes any nonzero ${}_nH_i \times 1$ vector, we obtain, from (3.4),

$$\widetilde{\lambda}_D^{K(i+1)}(t) = \det \widetilde{K}_D^{(i-1)} / \det \begin{pmatrix} \widetilde{K}_D^{(i-1)}, & \widetilde{P}^{(i-1)}v \\ v^*(\widetilde{P}^{(i-1)})^*, & v^* k_{ii} v \end{pmatrix} \tag{4.6}$$

$$= 1/kv^* \widetilde{\Omega}_D^{(i)}(t) v > 0$$

and hence $\widetilde{\Omega}_D^{(i)}(z)$ is positive definite and also $\det \widetilde{\Omega}_D^{(i)}(z) > 0$ follows. Therefore, we have, from (4.1) and (4.4),

$$\det \widetilde{K}_D^{(i)}(z, \bar{z}) = k \det \widetilde{K}_D^{(i-1)}(z, \bar{z}) \det \widetilde{\Omega}_D^{(i)}(z) > 0.$$

Under the condition $(f(t), D_z f(t), \ldots, D_z^{i-1} f(t), D_z D_z^{i-1} f(z)|_{z=t}(u \times v)) = (0, \ldots, 0, 1) \equiv K(i+1)$, where u and v are $n \times 1$ and ${}_nH_{i-1} \times 1$ constant vector respectively, we have, by the same procedure as above,

$$(u \times v)^* \Omega_D^{(i)}(t)(u \times v) = v^*(u^* \times E_{t(i-1)}) \Omega_D^{(i)}(t)(u \times E_{t(i-1)}) v > 0,$$

which shows (4.5).

Lemma 4.2. *In a bounded domain D, we have*

$$\begin{aligned} &\partial_z^* \partial_z \log \det \widetilde{K}_D^{(i)}(z, \bar{z}) \\ &= \mathrm{Sp}\{(\widetilde{\Omega}_D^{(i)})^{-1}(dz^* \times E_{t(i)}) \Omega_D^{(i+1)}(z)(dz \times E_{t(i)})\} > 0. \end{aligned} \tag{4.7}$$

Proof. Noting that $\widetilde{\Omega}_D^{(i)}(z)$ and $(u^* \times E_{t(i)}) \Omega_D^{(i+1)}(z)(u \times E_{t(i)})$ are positive definite from (4.4), we have by Lemma 2.1

$$\begin{aligned} &\partial_z^* \partial_z \log \det \widetilde{K}_D^{(i)} \\ &= \mathrm{Sp}[(\widetilde{K}_D^{(i)})^{-1}(dz^* \times E_{s(i)})\{\widetilde{K}_{D,11}^{(i)} - \widetilde{K}_{D,10}^{(i)}(\widetilde{K}_D^{(i)})^{-1}\widetilde{K}_{D,01}^{(i)}\}(dz \times E_{s(i)})] \\ &= \mathrm{Sp}\left\{\begin{pmatrix} * & * \\ * & (k\widetilde{\Omega}_D^{(i)})^{-1} \end{pmatrix} \begin{pmatrix} dz^* & 0 \\ 0 & dz^* \times E_{t(i)} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & k\Omega_D^{(i+1)} \end{pmatrix} \begin{pmatrix} dz & 0 \\ 0 & dz \times E_{t(i)} \end{pmatrix}\right\} \\ &= \mathrm{Sp}\{(\widetilde{\Omega}_D^{(i)})^{-1}(dz^* \times E_{t(i)}) \Omega_D^{(i+1)}(dz \times E_{t(i)})\} > 0, \end{aligned}$$

where $s(i) = 1 + {}_nH_1 + \cdots + {}_nH_i = \binom{n+i}{i}$ and $t(i) = {}_nH_i$, since $\mathrm{Sp}(H_1H_2) > 0$ follows when H_1 and H_2 are positive definite Hermitian matrices.

Definition 4.2. Such an ${}_nH_i \times {}_nH_j$ matrix $\sigma(A)$ that

$$([v]^i)^T A[u]^j = (v^i)^T \sigma(A) u^j$$

holds for arbitrary nonzero vectors $u \equiv (u_1, \ldots, u_n)^T$ and $v \equiv (v_1, \ldots, v_n)^T$ is called the σ-contraction of an $n^i \times n^i$ matrix A.

Further, for a linear transformation $v = Au$ we define another contraction $\delta[A]^k$ of $[A]^k$ as follows: $v^k = (\delta[A]^k)u^k$, where u, v and A denote $n \times 1$, $m \times 1$ vectors and an $m \times n$ matrix, respectively.

Lemma 4.3. *Let $g(z, \bar{z})$ and $w(z)$ be a scalar function and a biholomorphic mapping in D, then we have*

$$\delta[AB]^k = (\delta[A]^k)(\delta[B]^k), \tag{4.8}$$

in particular, $\delta[u]^k = u^k$ and $\delta[Au]^k = (\delta[A]^k)u^k$, and further we have

$$\begin{gathered} \sigma(g_{[ij]}) = g_{ij}, \\ \sigma\{([D_z w]^i)^* g_{[ij]} [D_z w]^j\} = \delta([D_z w]^i)^* g_{ij} \delta[D_z w]^j. \end{gathered} \tag{4.9}$$

For an $n \times n$ matrix C and a natural number k we have

$$\det \delta[C]^k = (\det C)^{s(k-1)}, \qquad s(k-1) = \binom{n+k-1}{k-1}. \tag{4.10}$$

Proof. (4.8) and (4.9) are evident from Definition 4.2.

By the triangulation of C we have $C = PSP^{-1}$, where P and S denote $n \times n$ regular and $n \times n$ triangular matrices, respectively. Since $[C]^k = [P]^k[S]^k[P^{-1}]^k$ and $\delta[P^{-1}]^k = \delta([P]^k)^{-1} = (\delta[P]^k)^{-1}$ hold, then we obtain $\det(\delta[C]^k - \lambda E_n) = \det(\delta[S]^k - \lambda E_n)$, which derives (4.10). $s(k-1)$ is obtained from $t(k) \times k/n = {}_nH_k k/n = {}_{n+1}H_{k-1} = \binom{n+k-1}{k-1}$.

Lemma 4.4. *Under $w(z) \in BH(D)$ we have the relative invariances:*

$$\det \widetilde{K}_D^{(i)}(z, \bar{z}) = \det \widetilde{K}_\Delta^{(i)}(w, \bar{w}) |J_w(z)|^{2N(i)}, \qquad i \geq 0, \tag{4.11}$$

and have the absolute invariants:

$$I_D^{(i)}(z) \equiv \det \widetilde{K}_D^{(i)}(z, \bar{z}) / (k_D(z, \bar{z}))^{N(i)}, \qquad i \geq 0, \tag{4.12}$$

where $\Delta = w(D)$ and $N(i) \equiv \binom{n+i+1}{i}$.

In particular, for $i = 1$ we have a known absolute invariant:

(4.13) $$I_D^{(1)}(z) \equiv \det T_D(z, \bar{z})/k_D(z, \bar{z}) \qquad [3].$$

Proof. The Bergman kernel function $k_D(z, \bar{z})$ has the relative invariancy:

(4.14) $$k_D(z, \bar{z}) = \bar{J} k_\Delta(w, \bar{w}) J \quad \text{for } w(z) \in BH(D),$$

where $J \equiv J_w(z) = \det(D_z w)$. Let us set $k_D(z, \bar{z}) \equiv k_D$ and $k_\Delta(w, \bar{w}) \equiv k_\Delta$; then we have

$$[D_z^*]^p [D_z]^q k_\Delta = ([D_z w]^p)^* ([D_w^*]^p [D_w]^q k_\Delta)[D_z w]^q.$$

Since

$$[D_z^*]^p [D_z]^q k_D = [D_z^*]^p [D_z]^q (\bar{J} k_\Delta J)$$

(4.15) $$= \sum_{j=0}^{p} \sum_{i=0}^{q} ({}_pC_j \times {}_qC_i) \{([D_z w]^{p-j})^* \times \bar{J}_{[j0]}\} \cdot k_{\Delta,[p-j,q-i]} \{[\dot{D}_z w]^{q-i} \times J_{[0i]}\},$$

using the elementary theorems with respect to the determinant and the contraction

(4.16) $$\sigma([D_z^*]^p [D_z]^q k_\Delta) = \delta([D_z w]^p)^* k_{\Delta,pq} \delta[D_z w]^q,$$

we have, by (4.10),

$$\det \widetilde{K}_D^{(i)}(z, \bar{z}) = \det \begin{pmatrix} \cdots & \cdots & \cdots \\ \cdots & (\bar{J} k_\Delta J)_{pq} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}$$

(4.17) $$\begin{aligned} &= |J|^{2\Sigma_{k=0}^{i} t(k)} \det \begin{pmatrix} \cdots & \cdots & \cdots \\ \cdots & \delta([D_z w]^p)^* k_{\Delta,pq} \delta[D_z w]^q & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \\ &= |J|^{2\Sigma_{k=0}^{i} t(k)} \left| \prod_{q=1}^{i} \det \delta[D_z w]^q \right|^2 \det \widetilde{K}_\Delta^{(i)}(w, \bar{w}) \\ &= |J|^{2(\Sigma_{k=0}^{i} t(k) + \Sigma_{k=0}^{i-1} s(k))} \det \widetilde{K}_\Delta^{(i)}(w, \bar{w}), \end{aligned}$$

where $t(k) = {}_nH_k$ and $s(k) = \binom{n+k}{k}$. Since

$$\sum_{k=0}^{i} t(k) + \sum_{k=0}^{i-1} s(k) = \binom{n+i}{i} + \binom{n+i}{i-1} = \binom{n+i+1}{i} \equiv N(i),$$

we have (4.11) and thus (4.12).

Theorem 4.1. *In a bounded domain D*

$$(4.18) \qquad (ds_D^{(i)})^2 \equiv \partial_z^* \partial_z \log \det \widetilde{K}_D^{(i)}(z, \bar{z}), \qquad i = 0, 1, 2, \cdots,$$

define the new invariant Kähler metrics under $BH(D)$ *(see* (4.7)).

Proof. The positivity of each $(ds_D^{(i)})^2$ is given by Lemma 4.2.

From Lemma 4.4 we can obtain the invariancy of $(ds_D^{(i)})^2$ under $BH(D)$, since we have

$$\log \det \widetilde{K}_D^{(i)}(z, \bar{z}) = \log \det \widetilde{K}_\Delta^{(i)}(w, \bar{w}) + \psi(z) + \overline{\psi(z)},$$

where $\psi(z)$ denotes the scalar analytic function $N(i) \log J_w(z)$, and $\partial_z f(z) = (D_w f(z(w))(D_z w))dz = (D_w F(w))dw$ holds for a holomorphic function $f(z) = f(z(w)) \equiv F(w)$ under $w(z) \in BH(D)$.

Remark 4.1. $(ds_D^{(0)})^2$ and $(ds_D^{(1)})^2$ coincide with the Bergman metric [3] and the Fuks metric [8], respectively.

Corollary 4.1. *In a bounded domain D, we have*

$$(4.19) \qquad -(R_{\bar{\alpha}\beta}) + (n+1)T_D = D_z^* D_z \log \det K_D^{(1)} \qquad (cf. [13])$$

and for any nonzero vector u

$$(4.20) \quad u^*(D_z^* D_z \log \det K_D^{(1)})u = \mathrm{Sp}\{T_D^{-1}(u^* \times E_n)\Omega_D^{(2)}(z)(u \times E_n)\} > 0,$$

where, for the Hermitian curvature tensor $(-R_{\bar{\alpha}\beta\bar{\gamma}\delta})$,

$$(4.21) \quad (R_{\bar{\alpha}\beta}) \equiv \left(\sum_{\gamma\delta} T^{\bar{\gamma}\delta}(-R_{\bar{\alpha}\beta\bar{\gamma}\delta})\right) = -D_z^* D_z \log \det T_D, \qquad (T^{\bar{\gamma}\delta}) = T_D^{-1},$$

[13] *denotes the Ricci tensor with respect to the Bergman metric.*

Proof. By Lemma 4.2 we have

$$\begin{aligned}\partial_z^* \partial_z \log \det K_D^{(1)} &= \partial_z^* \partial_z \log(k^{n+1} \det T) = dz^*(-(R_{\bar{\alpha}\beta}) + (n+1)T) \\ &= \mathrm{Sp}\{(\widetilde{\Omega}_D^{(1)}(z))^{-1}(dz^* \times E_n)\Omega_D^{(2)}(z)(dz \times E_n)\} \\ &= \mathrm{Sp}\{T^{-1}(dz^* \times E_n)\Omega_D^{(2)}(z)(dz \times E_n)\} > 0\end{aligned}$$

since $\widetilde{K}_D^{(1)} \equiv K_D^{(1)}$ holds, where $k \equiv k_D(z, \bar{z})$ and $T \equiv T_D(z, \bar{z})$.

Corollary 4.2. *In a bounded domain D, let us set*

$$J_{D,(p,q)}(z,\bar{z}) \equiv \det(k_D^p(z,\bar{z}) \times T_D^q(z,\bar{z})), \tag{4.22}$$

which is relatively invariant under $BH(D)$ for arbitrary real number p and integer q; then

$$ds^2_{D,(p,q)} \equiv \partial_z^* \partial_z \log J_{D,(p,q)}(z,\bar{z}) \quad (\equiv dz^* T_{D,(p,q)}(z,\bar{z})\, dz) \tag{4.23}$$

defines an invariant Kähler metric under $BH(D)$ for each (p, q) such that $np - (n+1)q \geq 0$ ($n = \dim D$). Here $k_D^p(z,\bar{z})$ takes values of the real positive branch.

Proof. Since $J_{D,(p,q)}(z,\bar{z}) = k_D^{pn}(z,\bar{z})(\det T_D(z,\bar{z}))^q$, then

$$\partial^2 \log J_{D,(p,q)}/\partial z^* \partial z = pnT_D - q(R_{\bar{\alpha}\beta}) > pnT_D - q(n+1)T_D \geq 0$$

follows from (4.18) and (4.19). The invariancy of $ds^2_{D,(p,q)}$ follows from the relative invariancies of k_D and T_D. We can obtain the relative invariance $J_{D,(p,q)}(z,\bar{z}) = J_{\Delta,(p,q)}(w,\bar{w})|J_w(z)|^{2(pn+q)}$ for $w(z) \in BH(D)$ and $\Delta = w(D)$, where $|J_w(z)|^{2(pn+q)}$ takes values of the real positive branch.

Remark 4.2. The particular case of $(p, q) = ((n+1)/n, 1)$ ($n = \dim D$) was treated by Fuks [8] and $ds^2_{D,((n+1)/n,1)}$ coincides with $(ds_D^{(2)})^2$. For $(p, q) = (2, 1)$, $ds^2_{D,(2,1)}$ coincides with the Kato metric [11], which is valid for arbitrary n ($n = \dim D$) and for $(p, q) = (1, 0)$, $ds^2_{D,(1,0)}$ denotes the Bergman metric.

Under the restriction $q = 1$ and $p \geq (n+1)/n$, (i) the possible minimum value of p for each n ($n = \dim D$) equals $(n+1)/n$, which is the case of Fuks, and (ii) the possible maximum value of p for all n ($n = \dim D \geq 1$) equals 2, which is the case of Kato.

If D is a bounded homogeneous domain, $ds^2_{D,(p,q)}$ is essentially equivalent to the Bergman metric for $pn + q > 0$.

Corollary 4.3. *In a bounded domain D, we have*

$$\Omega_D^{(2)}(z) = K_{[22,00]} - K_{[21,00]}T^{-1}K^*_{[21,00]}, \tag{4.24}$$

$$(u^* \times E_n)\Omega_D^{(2)}(z)(u \times E_n) > 0 \quad (\textit{positive definite}) \textit{ in } D \tag{4.25}$$

and further

$$\tilde{\Omega}_D^{(2)}(z) = K_{22,00} - K_{21,00}T^{-1}K^*_{21,00} > 0 \quad \textit{in } D. \tag{4.26}$$

Here, $K_{[ij,st]} = (k_{[ij]} \times k_{[st]} - k_{[it]} \times k_{[sj]})/k^2,$

$$K_{ij,st} = (k_{ij} \times k_{st} - k_{it} \times k_{sj})/k^2,$$

$T \equiv T_D(z, \bar{z})$ *and* $k \equiv k_D(z, \bar{z})$, *and* u *and* v *denote nonzero* $n \times 1$ *vectors.*

Proof. From (4.5) we have

$$\Omega_D^{(2)}(z) = \{k_{[22]} - (P^{(1)})^*(\tilde{K}_D^{(1)})^{-1}P^{(1)}\}/k^2. \quad \square$$

Noting (3.9), we have (4.24) by straight calculations.

It is known [3] that the Hermitian curvature tensor $(-R_{\bar{\alpha}\beta\bar{\gamma}\delta})$ of the first kind with respect to the Bergman metric $ds^2 = dz^*T_D(z, \bar{z})dz$ of D is given by

$$\begin{aligned}(-R_{\bar{\alpha}\beta\bar{\gamma}\delta}) &= -T_{2,D}(z, \bar{z}) \\ &\equiv -(T_{11} - T_{10}T^{-1}T_{01}) = -(E_n \times T)D_z^*(T^{-1}D_zT),\end{aligned} \tag{4.27}$$

where $T \equiv T_D(z, \bar{z})$ and $T^{-1}D_zT$ denotes the matrix of the Christoffel symbols.

Theorem 4.2. *The Hermitian curvature tensor with respect to the Bergman metric has the following expression:*

$$-T_{2,D}(z, \bar{z}) = (T_D(z, \bar{z}) \times T_D(z, \bar{z}))(E_n \times E_n + \tilde{E}_{nn}) - \Omega_D^{(2)}(z) \tag{4.28}$$

(*cf.* [13]).

$\Omega_D^{(2)}(z)$ *is a relative invariant under* $BH(D)$.

Proof. Noting that $k^*_{[ij]} = k_{[ji]}$, $k_{[i0]} \times k_{[0j]} = k_{[0j]} \times k_{[i0]}$, $T^*_{01} = T_{10}$ and $D_z^*H = (D_zH)^*$ for an Hermite matrix $H(z, \bar{z})$, we have, by differentiating both sides of $k^2 \times T = k \times k_{11} - k_{10} \times k_{01}$ with respect to z and z^*,

$$\begin{aligned}k^2 \times (D_z^*D_z(k^2 \times T)) - (D_z^*(k^2 \times T))T^{-1}(D_z(k^2 \times T)) &= k^4 \times (T_{2,D} + 2T \times T) \\ &= k^2 \times (k \times k_{[22]} - k_{[20]} \times k_{[02]}) \\ &\quad - (k \times k_{[21]} - k_{[20]} \times k_{01})T^{-1}(k \times k_{[12]} - k_{10} \times k_{[02]}) \\ &\quad + k^2 \times k_{11} \times k_{11}(E_n \times E_n - \tilde{E}_{nn}) \\ &\quad - (k_{10} \times k_{11} - k_{11} \times k_{10})T^{-1}(k_{01} \times k_{11} - k_{11} \times k_{01}),\end{aligned}$$

since $k_{01} \times k_{11} - k_{11} \times k_{01} = k \times (k_{01} \times T - T \times k_{01})$. Noting that

$$(k_{10} \times T) \times k_{01} = \{T \times (k_{10} \times k_{01})\}\widetilde{E}_{nn},$$

$$(k_{10} \times k_{01}) \times (k_{10} \times k_{01}) = \{(k_{10} \times k_{01}) \times (k_{10} \times k_{01})\}\widetilde{E}_{nn}$$

and

$$k_{11}/k = T + k_{10} \times k_{01}/k^2,$$

we obtain

$$\begin{aligned}(k_{11} \times k_{11})&(E_n \times E_n - \widetilde{E}_{nn})/k^2 \\ &- (k_{10} \times k_{11} - k_{11} \times k_{10})T^{-1}(k_{01} \times k_{11} - k_{11} \times k_{01})/k^4 \\ = (T \times T)&(E_n \times E_n - \widetilde{E}_{nn}).\end{aligned}$$

Thus we get (4.28).

Since $T \times T$ and $T_{2,D}$ are relatively invariant under $BH(D)$ [10], [13], [14] and $[D_z w]^2 \widetilde{E}_{nn} = \widetilde{E}_{nn}[D_z w]^2$ holds, then it follows from (4.28) that $\Omega_D^{(2)}(z)$ is relatively invariant under $BH(D)$.

Theorem 4.3. *For each i $(i = 0, 1, 2, \dots)$ the mapping*

$$w_D^{(i)}(z) \equiv T_D^{(i)}(t, \bar{t}) \int_t^z T_D^{(i)}(z, \bar{t})\, dz + t, \qquad t \in D, \tag{4.29}$$

defines the ith representative function, i.e., any domain Δ in the equivalent class $F \equiv \{f(D) | f(z) \in BH(D),\ f(t) = t,\ D_z f(t) = E_n\}$ is mapped onto the (unique) ith representative domain with center at t by the function $w = w_\Delta^{(i)}(z)$, where $T_D^{(i)}(z, \bar{z})$ denotes the fundamental tensor $D_z^* D_z \log\det \widetilde{K}_D^{(i)}(z, \bar{z})$ for the ith metric (4.18).

A bounded domain D is an ith representative domain with center at t if and only if

$$T_D^{(i)}(z, \bar{t}) = T_D^{(i)}(t, \bar{t}) \quad \text{in } D \tag{4.30}$$

holds (*see* [17]).

Proof. Since $T_D^{(i)}(z, \bar{t})$ is relatively invariant under $BH(D)$, then we have $w = w_D^{(i)}(z) = w_\Delta^{(i)}(\zeta)$ under any $\zeta = \zeta(z) \in F$. The latter half of the theorem is easily obtained by $w_D^{(i)}(z) \equiv z$ in D.

5. **Curvatures and estimations.** For the general sectional curvature

$R_D(z;\, u, v, u, v)$ (which is the expression in differential geometry) and a complex structure J, the holomorphic bisectional curvature with respect to the Bergman metric is defined as $R_D(z;\, u, Ju, v, Jv)$ (S. Kobayashi). After some direct calculations we can show that $R_D(z;\, u, Ju, v, Jv)$ coincides with the unitary curvature $R_D(z;\, u, v)$ due to Hua [10] (see (4.27)). Now, we shall give the matrix expressions of the holomorphic bisectional curvature $R_D(z;\, u, v)$ (of course, $R_D(z;\, u, u)$ coincides with the holomorphic sectional curvature $R_D(z;\, u)$), the Ricci curvature

$$C_D(z;\, u) \equiv u^*(R_{\overline{\alpha}\beta})u/u^*T_Du$$

and the Ricci scalar curvature

$$S_D(z) \equiv \mathrm{Sp}\{T_D^{-1}(R_{\overline{\alpha}\beta})\} = \sum_{\overline{\alpha}\beta\overline{\gamma}\delta} T^{\overline{\alpha}\beta}T^{\overline{\gamma}\delta}(-R_{\overline{\alpha}\beta\overline{\gamma}\delta})$$

in terms of $T \equiv T_D(z, \overline{z})$ (Bergman metric tensor) and $\dot{T}_{2,D} \equiv T_{2,D}(z, \overline{z})$ (see (4.27) and (4.28)).

Lemma 5.1. *For a bounded domain D in C^n and contravariant section vectors u and v, we have*

$$R_D(z;\, u, v) = -(u \times v)^*T_{2,D}(u \times v)/u^*Tuv^*Tv, \tag{5.1}$$

$$C_D(z;\, u) = -\mathrm{Sp}\{T^{-1}(u^* \times E_n)T_{2,D}(u \times E_n)\}/u^*Tu \tag{5.2}$$

and

$$S_D(z) = -\mathrm{Sp}\{(T^{-1} \times T^{-1})T_{2,D}\}, \tag{5.3}$$

which are absolute invariants under $BH(D)$.

Proof. Using the formula (2.5) and (4.19), we obtain

$$\begin{aligned} C_D(z;\, u) &= -u^*(D_z^*D_z \log\det T)u/u^*Tu \\ &= -\mathrm{Sp}\{T^{-1}(u^* \times E_n)T_{2,D}(u \times E_n)\}/u^*Tu. \end{aligned}$$

$S_D(z) \equiv \Sigma\, T^{\overline{\alpha}\beta}T^{\overline{\gamma}\delta}(-R_{\overline{\alpha}\beta\overline{\gamma}\delta}) = -\mathrm{Sp}\{(T^{-1} \times T^{-1})T_{2,D}\}$ is evident.

The biholomorphic invariancies of (5.1), (5.2) and (5.3) are easily obtained by the relative invariancies of T and $T_{2,D}$ under $BH(D)$ [14]. □

For an $n \times n$ matrix $B = (b_{ij})$ and $n \times 1$ vectors

$$M_i \equiv (0, \ldots, 0, 1, 0, \ldots, 0)^T,$$

where 1 occurs in the ith position ($i = 1, \ldots, n$), we have

$$(5.4) \qquad M_j^T B M_i = b_{ji}, \qquad \sum_{i=1}^{n} M_i^T B M_i = \mathrm{Sp}(B).$$

Lemma 5.2. *Let* v_i *be the mutually orthogonal sections* $T^{-1/2}M_i$ ($i = 1, \ldots, n$) *such that* $v_j^* T v_i = M_j^T M_i = \delta_{ij}$; *then we have*

$$(5.5) \qquad C_D(z; u) = \sum_{i=1}^{n} R_D(z; u, v_i)$$

and

$$(5.6) \qquad S_D(z) = \sum_{j=1}^{n} C_D(z; v_j) = \sum_{i,j=1}^{n} R_D(z; v_j, v_i).$$

Proof. From (5.1) for $v = v_i$, noting $v_i^* T v_i = 1$, we have

$$R_D(z; u, v_i) = -M_i^T T^{-1/2}(u^* \times E_n)T_{2,D}(u \times E_n)T^{-1/2}M_i/u^*Tu.$$

By summation with respect to i we obtain, from (5.2),

$$\sum_{i=1}^{n} R_D(z; u, v_i) = -\mathrm{Sp}\{T^{-1}(u \times E_n)T_{2,D}(u \times E_n)\}/u^*Tu = C_D(z; u).$$

By the same procedure, we have

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{n} R_D(z; v_j, v_i) &= \sum_{j=1}^{n} C_D(z; v_j)\\ &= -\mathrm{Sp}\{(T^{-1/2} \times T^{-1/2})T_{2,D}(T^{-1/2} \times T^{-1/2})\}\\ &= -\mathrm{Sp}\{(T^{-1} \times T^{-1})T_{2,D}\} = S_D(z).\end{aligned}$$

Theorem 5.1. *Let* $\lambda^{(1)}$, $\lambda^{(2)}(u)$ *and* $\lambda^{(3)}(u, v)$ *be the minimum values in* (3.12), (3.11) *and* (3.13) *at* z *in a bounded domain* D, *respectively, and* $\epsilon \equiv \epsilon_D(z; u, v)$ *be* $|u^*Tv|^2/u^*Tuv^*Tv$ ($0 \le \epsilon \le 1$ *for* $n \ge 2$ *and* $\epsilon = 1$ *for* $n = 1$ *and* $\epsilon_D(z; u, u) = 1$); *then we have, for any sections* u *and* v,

$$\begin{aligned}R_D(z; u,v) &= 1 + \epsilon - (u \times v)^*\Omega(u \times v)/u^*Tuv^*Tv\\ (5.7) \qquad &= 1 + \epsilon - \lambda^{(2)}(u)\lambda^{(2)}(v)/\lambda^{(1)}\lambda^{(3)}(u, v) < 2 \quad (cf.\ [2], [6], [19]),\end{aligned}$$

$$(5.8)\quad \begin{aligned} C_D(z;\, u) &= n + 1 - \mathrm{Sp}\{T^{-1}(u^* \times E_n)\Omega(u \times E_n)\}/u^*Tu \\ &= n + 1 - \lambda^{(2)}(u) \sum_{i=1}^{n} (\lambda^{(3)}(u,\, v_i))^{-1} < n + 1 \quad (cf.\ [4]) \end{aligned}$$

and

$$(5.9)\quad \begin{aligned} S_D(z) &= n(n + 1) - \mathrm{Sp}\{(T^{-1} \times T^{-1})\Omega\} \\ &= n(n + 1) - \lambda^{(1)} \sum_{i,j=1}^{n} (\lambda^{(3)}(v_i,\, v_j))^{-1} < n(n + 1), \end{aligned}$$

where $\Omega \equiv \Omega_D^{(2)}(z)$ *and* $v_i \equiv T^{-1/2}M_i$ $(i = 1, \ldots, n)$ *are given in* (4.4) *and Lemma* 5.2, *respectively.*

Proof. By (5.1), (4.28) and Lemma 3.1 we have (5.7).

Since it follows from (3.11) that $\lambda^{(2)}(v_i) = \lambda^{(1)}$, then we have

$$\begin{aligned} C_D(z;\, u) &= \sum R_D(z;\, u,\, v_i) \\ &= n + \left[\sum |u^*T^{1/2}M_i|^2 - \mathrm{Sp}\{T^{-1}(u^* \times E_n)\Omega(u \times E_n)\}\right] \Big/ u^*Tu \\ &= n + 1 - \mathrm{Sp}\{T^{-1}(u^* \times E_n)\Omega(u \times E_n)\}/u^*Tu \end{aligned}$$

for any section vector $u = \Sigma_{j=1}^n b_j v_j$ $(\Sigma_{j=1}^n |b_j|^2 = 1)$. (5.9) follows from (5.8) and (5.6).

Remark 5.1. $R_D(z;\, u) \equiv R_D(z;\, u,\, u) = 2 - (\lambda^{(2)}(u))^2/\lambda^{(1)}\lambda^{(3)}(u,\, u) < 2$ [2] and $R_D(z;\, u,\, v) < 2$ [10] are known.

Let u_0 and v_0 be any orthogonal vectors such as $u_0^*Tv_0 = 0$; then we have, for $n \geq 2$,

$$(5.10)\qquad R_D(z;\, u_0,\, v_0) < 1.$$

In a bounded homogeneous domain D, the absolute invariant $I_D^{(1)}(z,\, \bar{z})$ under $BH(D)$ (see (4.13)) equals a positive constant in D. Therefore, a domain D with $I_D^{(1)} \equiv$ constant or a homogeneous domain D satisfies, for any section vector u,

$$(5.11)\qquad C_D(z;\, u) = -1 \quad \text{and} \quad S_D(z) = -n \qquad \text{in } D.$$

Let G be a bounded domain in C^1, then we easily have $R_G(z;\, u,\, v) = R_G(z;\, u) = C_G(z;\, u) = S_G(z)$. If G is also homogeneous, we have $R_G(z;\, u,\, v) = -1$ in G since G is symmetric by Cartan's theorem and hence is simply connected.

Theorem 5.2. *Let D be a bounded domain in C^n $(n \geq 2)$; then we have, for any section vectors u and v,*

$$(5.12) \qquad -n + \epsilon + C_D(z; u) < R_D(z; u, v) < 1 + \epsilon \quad in\ D.$$

In particular, if D is homogeneous, then we have, for any section vectors u and v,

$$(5.13) \qquad -(n+1) + \epsilon < R_D(z; u, v) < 1 + \epsilon \quad in\ D,$$

$$(5.14) \qquad -n < R_D(z; u) < 2 \quad in\ D\ (cf.\ [10])$$

and there exist some vectors u' and v' such that

$$(5.15) \qquad R_D(z; u', v') < 0 \quad in\ D.$$

Proof. Let A be a positive definite Hermitian $n \times n$ matrix and $v = T^{-1/2}P$ be a vector with $P^*P = 1$ i.e., v denotes a vector $\Sigma_{i=1}^n p_i v_i$, where $P \equiv (p_1, \ldots, p_n)^T$ and $v_i = T^{-1/2}M_i$ (see (5.4)), then we have $v^*Av \leq \mathrm{Sp}(T^{-1}A)$ (inequality for $n \geq 2$ and equality for $n = 1$). For any vector v with $v^*Tv = 1$, we have (5.12) from (5.7) and (5.8), since we have

$$(u \times v)^*\Omega(u \times v)/u^*Tuv^*Tv < \mathrm{Sp}\{T^{-1}(u^* \times E_n)\Omega(u \times E_n)\}/u^*Tu$$

from (4.25).

If D is homogeneous, we have (5.13) from (5.12) and (5.11). (5.14) is easily obtained by $\epsilon_D(z; u, u) = 1$ in D. From (5.5) and (5.11) we have

$$n\left\{\inf_{u,v} R_D(z; u, v)\right\} \leq -1 \leq n\left\{\sup_{u,v} R_D(z; u, v)\right\}$$

and hence (5.15).

Theorem 5.3. *Let D be a bounded homogeneous domain and $(u^* \times E_n)T_{2,D}(u \times E_n)$ be nonnegative definite (resp. positive definite); then we have, for $n \geq 2$,*

$$(5.16) \qquad -1 \leq R_D(z; u, v) \leq 0 \qquad (resp.\ -1 < R_D(z; u, v) < 0).$$

Proof. For any section vector $v = T^{-1/2}P$ with $P^*P = 1$, we have $R_D(z; u, v) = -P^*QP/u^*Tu$, where $Q \equiv T^{-1/2}(u^* \times E_n)T_{2,D}(u \times E_n)T^{-1/2} = U^*(\lambda_1 \dotplus \cdots \dotplus \lambda_n)U$ (U: unitary $n \times n$ matrix and $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$), since $T \equiv T_D(z, \bar{z})$ and $(u^* \times E_n)T_{2,D}(u \times E_n)$ are positive and nonnegative (resp. positive) definite, respectively. Set $UP = S \equiv (s_1, \ldots, s_n)^T$, then we have $S^*S = 1$. Let D be a homogeneous domain with $Q \geq 0$, then it fol-

lows that $-1 \leq S_D(z;\, u) = -\mathrm{Sp}(Q)/u^*Tu = -\Sigma_{i=1}^n \lambda_i/u^*Tu$ and thus $\Sigma_{i=1}^n \lambda_i = u^*Tu > 0$, i.e., $\lambda_1 > 0$. Hence we get

$$-1 \leq -\lambda_1 \Big/ \sum_{i=1}^n \lambda_i \leq R_D(z;\, u,\, v)$$

$$= -\sum_{i=1}^n \lambda_i |s_i|^2 \Big/ \sum_{i=1}^n \lambda_i \leq -\lambda_n \Big/ \sum_{i=1}^n \lambda_i \leq 0.$$

Example 5.1. Any classical Cartan domain D satisfies that $v^*(u^* \times E_n)T_{2,D}(u \times E_n)v \geq 0$ for any section vector v. Therefore, (5.16) holds in D. Let $R(i)$ ($i = \mathrm{I, II, III, IV}$) be the classical Cartan domains (four main types of irreducible bounded symmetric domains). They are homogeneous, and the following hold [14]:

$$-2/(m+n) \leq R_{R(\mathrm{I})}(z;\, u) \leq -2/m(m+n) \qquad (m \geq n \geq 1),$$

$$-2/(n+1) \leq R_{R(\mathrm{II})}(z;\, u) \leq -2/n(n+1),$$

$$-1/(n-1) \leq R_{R(\mathrm{III})}(z;\, u) \leq -1/[n/2](n-1) \qquad (n \geq 2),$$

$$-2/n \leq R_{R(\mathrm{IV})}(z;\, u) \leq -1/n.$$

For the n-polydisc P and the unit hypersphere E

(5.17) $\quad -1 \leq R_P(z;\, u) \leq -1/n \quad$ and $\quad R_E(z;\, u) = -2/(n+1) \quad$ in D

hold, but in general $R_E(z;\, u,\, v)$ is "not constant" for arbitrary vectors u and v.

6. **Domains of comparison.** The basic tool used here and in the next section is the so-called method of minimum integral [3] or the principle of minimum problems [7].

Principle. Let $\lambda_A^{K(m)}(t)$ and $\lambda_B^{K(m)}(t)$ be the minimum values defined in §3 for two domains A and B with $A \subset B$ under the same additional condition $K(m)$ at $t \in A$; then we have

(6.1) $$\lambda_A^{K(m)}(t) \leq \lambda_B^{K(m)}(t).$$

Theorem 6.1. *Let A and B be domains of comparison of a bounded domain D ($A \subset D \subset B$) and $\epsilon_D(u,\, v) \equiv \epsilon_D(z;\, u,\, v)$; then we have, for $z \in A$,*

$$(1 + \epsilon_B(u, v) - R_B(z; u, v))/\Lambda_{AB}(u, v) \le 1 + \epsilon_D(u, v) - R_D(z; u, v)$$
$$\le (1 + \epsilon_A(u, v) - R_A(z; u, v))\Lambda_{AB}(u, v), \tag{6.2}$$

$$(n + 1 - C_B(z; u))/\Lambda_{AB}^{1/2}(u, u) \le n + 1 - C_D(z; u)$$
$$\le (n + 1 - C_A(z; u))\Lambda_{AB}^{1/2}(u, u) \tag{6.3}$$

and

$$(n(n + 1) - S_B(z))/\Psi_{AB} \le n(n + 1) - S_D(z) \le (n(n + 1) - S_A(z))\Psi_{AB}, \tag{6.4}$$

where

$$\Lambda_{AB}(u, v) \equiv \lambda_B^{(2)}(u)\lambda_B^{(2)}(v)/\lambda_A^{(2)}(u)\lambda_A^{(2)}(v)$$
$$= k_A^2 u^* T_A u v^* T_A v / k_B^2 u^* T_B u v^* T_B v$$

and

$$\Psi_{AB} \equiv \lambda_B^{(1)}/\lambda_A^{(1)} = k_A/k_B.$$

Proof. By Theorem 5.1 and Principle we have

$$1 + \epsilon_D(u, v) - R_D(z; u, v) = \lambda_D^{(2)}(u)\lambda_D^{(2)}(v)/\lambda_D^{(1)}\lambda_D^{(3)}(u, v)$$
$$\le (1 + \epsilon_A(u, v) - R_A(z; u, v))\Lambda_{AB}(u, v),$$

etc. Thus we have (6.2), (6.3) and (6.4) by the same procedure. □

By Theorem 6.1 and the biholomorphic invariancies of curvatures, we have the following:

Corollary 6.1. (i) *If A and B are image domains of the unit hypersphere and $A \subset D \subset B$ holds, then we have, for $z \in A$,*

$$2(1 - \nu\Lambda_{AB}(u, u)) \le R_D(z; u) \le 2(1 - \nu/\Lambda_{AB}(u, u)). \tag{6.5}$$

(ii) *If A and B are homogeneous domains of comparison of a bounded domain D, then we have, for $z \in A$,*

$$(n + 1)(1 - \nu\Lambda_{AB}^{1/2}(u, u)) \le C_D(z; u) \le (n + 1)(1 - \nu/\Lambda_{AB}^{1/2}(u, u)) \tag{6.6}$$

and

$$n(n + 1)(1 - \nu\Psi_{AB}) \le S_D(z) \le n(n + 1)(1 - \nu/\Psi_{AB}). \tag{6.7}$$

Here and in the following, ν denotes $(n + 2)/(n + 1)$.

Corollary 6.2. *If A and B are hyperspheres of radii r and R $(r < R)$ with the same center at the origin, respectively, and D $(A \subset D \subset B)$ is a homogeneous domain, then we have, for any section vector u and $x \in D$,*

$$(6.8) \qquad 2(1 - \nu(R/r)^{4n+4}) \leq R_D(x; u) \leq 2(1 - \nu(r/R)^{4n+4}).$$

Proof. For such a homogeneous mapping $h(z)$ of D that $h(t) = 0$ holds for any fixed point $t \in D$, we have $R_D(t; u) = R_D(0; v)$, where $v = D_z h(t)u$. On the other hand, from (6.5) we have

$$2(1 - \nu\Lambda_{AB}(v, v)) \leq R_D(0; v) \leq 2(1 - \nu/\Lambda_{AB}(v, v)).$$

The Bergman kernel function $k_A(z, \bar{z})$ and the Bergman metric tensor $T_A(z, \bar{z})$ of a hypersphere $A \equiv \{z \mid |z| < r, z \equiv (z_1, \ldots, z_n)^T\}$ are given by

$$(6.9) \qquad k_A(z, \bar{z}) = n!r^2/\pi^n(r^2 - z^*z)^{n+1}$$

and

$$(6.10) \qquad T_A(z, \bar{z}) = (n+1)r^2(r^2 \times E_n - zz^*)^{-1}/(r^2 - z^*z)$$

as is well known (see [14], [16]). Therefore, we have

$$\lambda_A^{(2)}(0; v) = 1/k_A(0, 0)v^*T_A(0, 0)v = \pi^n r^{2n+2}/n!v^*v$$

and hence $\Lambda_{AB}(u, u) = ((R/r)^{2n+2})^2$. Thus we obtain (6.5).

Remark 6.1. The holomorphic sectional curvatures of the classical Cartan domains are always negative as was stated before. All bounded symmetric domains are homogeneous but the converse is not true for $n \geq 4$ (E. Cartan). K. H. Look gave an example of a homogeneous but nonsymmetric domain D having a section u such that $R_D(z; u)$ has a positive value, which is the negative solution on the Hua's conjecture. For any homogeneous domain D, which satisfies $A \subset D \subset B$ and $(R/r)^{4n+4} \leq \nu$ in Corollary 6.2, we have $R_D(z; u) \leq 0$ for any u in D.

7. Asymptotic boundary behaviors of curvatures. Now, we shall study the behaviors of curvatures about a boundary point of a bounded domain D with a sort of convexity in C^n using the domains of comparison of D.

Definition 7.1. Let D be a domain in C^n. Suppose that there exists an analytic change of coordinates, one-to-one in a neighborhood Γ $(\Gamma \supset D)$ of a boundary point $P \in \partial D$, so that, with respect to this change of coordinates, $D \to \Delta$, $P \to Q = \{0\}$ $(Q \in \partial\Delta)$ and

$$\Delta = \{z \mid z_1 + \bar{z}_1 > z^*z + o(z^*z)\} \tag{7.1}$$

in the neighborhood of $Q = \{0\}$. Then Δ and also the original domain D are said to be strictly pseudoconvex globally representable (simply SPCGR) at Q and also at P, respectively. We call the new coordinates "normal" coordinates and the analytic hypersurface $z_1 = 0$ (with respect to the normal coordinates) is called the normal analytic hypersurface (simply NAH) [4], [9].

If $D \equiv \{z \mid \phi(z, \bar{z}) < 0$, $\phi \in C^2$-class in a neighborhood of $\bar{D}$, grad$(\phi) \neq 0$ on $\partial D\}$ in C^n is a strictly pseudoconvex domain in the sense of Levi at a point $Q = \{0\} \in \partial D$, i.e., ϕ satisfies $L(\phi(Q)) \equiv z^*(\partial^2\phi(Q)/\partial z^*\partial z)z > 0$ when $(\partial\phi(Q)/\partial z)z = 0$ and $z \neq 0$, then by the Taylor's expansion of ϕ at $Q = \{0\}$ and by suitable changes of coordinates (properly affine in C^n and biholomorphic in a neighborhood of $\bar{D}$), we have the image domain of the type of (7.1) (see [9, Theorem 3.5.1 and its proof]). Therefore, any strictly pseudoconvex domain (in the sense of Levi) with one-to-one "normal" analytic change of coordinates is a SPCGR domain. If D is a SPCGR domain, for the sake of estimates on curvatures, we can use Δ in (7.1) instead of D from the beginning, since curvatures are biholomorphically invariant.

The hypersphere

$$R_\delta \equiv \{\zeta \mid \zeta_1 + \bar{\zeta}_1 > \zeta^*\zeta + \delta\zeta^*\zeta,\ \delta\ (-1 < \delta < 1)\text{: real constant number}\} \tag{7.2}$$

is biholomorphically equivalent to the unit hypersphere $E \equiv \{z \mid |z| < 1\}$ under the transformation

$$T_\delta\colon z = (1 + \delta)\zeta - (1, 0, \ldots, 0)^T. \tag{7.3}$$

B. L. Chalmers [4] has given the domains of comparison $R_{-\epsilon}^{\alpha\beta}$ and $R_{\epsilon}^{\alpha'\beta'}$ $(\epsilon > 0)$ for a strictly (p, q) pseudoconvex globally representable domain D with the normal analytic hypersurface $h \equiv \{\zeta \mid \zeta_1 = 0\}$ lying entirely outside D. In the following, we shall treat a strictly $(1, n)$ pseudoconvex (usual pseudoconvex) globally representable domain (7.1) with the normal analytic hypersurface h lying entirely outside itself, which is called a SPCGR-NAH domain at Q.

$R_{-\epsilon}^{\alpha\beta}$ and $R_{\epsilon}^{\alpha'\beta'}$ are equivalent to the hypersphere $R_{-\epsilon}$ and R_ϵ (see (7.2)) under biholomorphic mappings

$$W\colon z_1 = \zeta_1/(1 - \alpha\zeta_1),\ z_k = \zeta_k\{1 + (\beta - \alpha)\zeta_1\}/(1 - \alpha\zeta_1), \qquad k = 2, \ldots, n, \tag{7.4}$$

and

$$W': z_1 = \zeta_1/(1+\alpha'\zeta_1),\ z_k = \zeta_k(1+\alpha'\zeta_1)/\{1+(\alpha'+\beta')\zeta_1\}, \tag{7.5}$$
$$k = 2, \ldots, n,$$

respectively. In particular, for sufficiently large numbers α, β, α' and β', we have

$$R_{\epsilon}^{\alpha'\beta'} \subset \Delta \subset R_{-\epsilon}^{\alpha\beta}, \tag{7.6}$$

where Δ denotes a SPCGR-NAH domain at $Q = \{0\}$ [4].

Definition 7.2. We shall write $\lim^A_{\zeta\to 0}$, or sometimes simply $\lim^A$, to indicate a limit is being taken as $\zeta \to 0$ in the set $0 < a < \mathrm{Re}(\zeta_1)/|\zeta|$ (a: positive constant number) and say $\zeta \to 0$ via an A-approach after Chalmers [4].

Lemma 7.1. *For a hypersphere* R_δ $(0 < \delta < 1)$ *we have*

$$\lim_{\zeta\to 0}{}^A (\zeta_1 + \bar{\zeta}_1)^{n+1} k_{R_\delta}(\zeta, \bar{\zeta}) = n!(1+\delta)^{n-1}/\pi^n \tag{7.7}$$

and for any constant nonzero vector $u \equiv (u_1, \ldots, u_n)^T$

$$\lim_{\zeta\to 0}{}^A (\zeta_1 + \bar{\zeta}_1)^2 u^* T_{R_\delta}(\zeta, \bar{\zeta}) u = \begin{cases} (n+1)|u_1|^2 & \text{for } u_1 \neq 0, \\ (n+1)(1+\delta)|u|^2 & \text{for } u_1 = 0. \end{cases} \tag{7.8}$$

Proof. Let E be a unit disc in C^n. Since $k_E(z, \bar{z}) = n!/\pi^n(1 - z^*z)^{n+1}$ (6.9) and $k_{R_\delta}(\zeta, \bar{\zeta}) = k_E(z, \bar{z})|J_z(\zeta)|^2 = k_E(z, \bar{z})(1+\delta)^{2n}$ for (7.3); then we have

$$k_{R_\delta}(\zeta, \bar{\zeta}) = n!(1+\delta)^{n-1}/\pi^n \Delta_\delta^{n+1},$$

where $1 - z^*z = (1+\delta)\Delta_\delta$ and $\Delta_\delta = \zeta_1 + \bar{\zeta}_1 - (1+\delta)|\zeta|^2$. Noting that $\lim^A_{\zeta\to 0}\Delta_{-\epsilon}/\Delta_\epsilon = 1$, we obtain (7.7).

Let us set $z \equiv U(z)\ (\rho, 0, \ldots, 0)^T$, where $U(z) = U(z(\zeta))$ denotes a unitary matrix and ρ $(\rho > 0) \to 1$ (for $z \to (-1, 0, \ldots, 0)^T$) is equivalent to $\zeta \to 0$ under (7.3). If we set $U^*(z)u = U^*(z(\zeta))u \equiv v \equiv (v_1, \ldots, v_n)^T$ and $\lim^A v = v_0 \equiv (v_1^0, \ldots, v_n^0)^T$, then we have $|v| = |v_0| = |u|$ and $v_1^0 = -u_1$, because $z^*u = (\rho, 0, \ldots, 0)U^*(z)u = \rho v_1 \to v_1^0$ and

$$z^*u = \{(1+\delta)\zeta^* - (1, 0, \ldots. 0)\}u = (1+\delta)\zeta^* u - u_1 \to -u_1$$

for an A-approach. Further, we have, from (6.10) and $T_{R_\delta}(\zeta, \bar{\zeta}) = (D_\zeta z)^* T_E(z, \bar{z}) D_\zeta z$,

$$T_{R_\delta}(\zeta, \bar{\zeta}) = (n+1)U(z)\{1 \dotplus (1+\delta)\Delta_\delta \dotplus \cdots \dotplus (1+\delta)\Delta_\delta\} U^*(z)/\Delta_\delta^2$$

and thus

$$u^* T_{R_\delta}(\zeta, \bar{\zeta}) u = (n+1) P_\delta / \Delta_\delta^2, \qquad P_\delta = |v_1|^2 + (1+\delta)\Delta_\delta \sum_{i=2}^{n} |v_i|^2.$$

Since we easily have $\lim^A_{\zeta \to 0} P_\delta = |u_1|^2$ and thus (7.8) for $u_1 \neq 0$.

If $u_1 = 0$, we have

$$\lim{}^A P_\delta / (\zeta_1 + \bar{\zeta}_1) = (1+\delta) \sum_{i=2}^{n} |v_i^0|^2 = (1+\delta)|u|^2,$$

because we have $z^* u = \rho v_1 = (1+\delta)\zeta^* u - u_1 = (1+\delta)\Sigma_{i=2}^n \bar{\zeta}_i u_i$ for $u_1 = 0$, and hence

$$P_\delta = \left| (1+\delta) \sum_{i=2}^{n} \bar{\zeta}_i u_i / \rho \right|^2 + (1+\delta)(\zeta_1 + \bar{\zeta}_1 - (1+\delta)|\zeta|^2) \sum_{i=2}^{n} |v_i|^2$$

$$= (1+\delta)(\zeta_1 + \bar{\zeta}_1) \sum_{i=2}^{n} |v_i|^2 + o(\zeta_1 + \bar{\zeta}_1)$$

follows from

$$\left| (1+\delta) \sum_{i=2}^{n} \bar{\zeta}_i u_i / \rho \right|^2 \Big/ (\zeta_1 + \bar{\zeta}_1) \leq (1+\delta)^2 |u|^2 |\zeta|^2 / \rho^2 (\zeta_1 + \bar{\zeta}_1) \to 0$$

and $(1+\delta)^2 |\zeta|^2 \Sigma_{i=2}^n |v_i|^2 / (\zeta_1 + \bar{\zeta}_1) \to 0$ for an A-approach. Now, noting (7.7), $\lim^A \Delta_{-\epsilon}/\Delta_\epsilon = 1$ and $\lim^A \Delta_\delta / (\zeta_1 + \bar{\zeta}_1) = 1$, we obtain (7.8) for $u_1 = 0$.

Lemma 7.2. *Setting* $R_\epsilon^{\alpha'\beta'} = A$ *and* $R_{-\epsilon}^{\alpha\beta} = B$, *we have*

$$\lim_{\zeta \to 0}{}^A \Psi_{AB}(\zeta, \bar{\zeta}) = \lim_{\zeta \to 0}{}^A k_A(\zeta, \bar{\zeta}) / k_B(\zeta, \bar{\zeta}) = \{(1+\epsilon)/(1-\epsilon)\}^{n-1} \tag{7.9}$$

and for any constant nonzero vector $u \equiv (u_1, \ldots, u_n)^T$

$$\lim_{\zeta\to 0}{}^{A}\Lambda_{AB}^{1/2}(\zeta,\bar\zeta) = \lim_{\zeta\to 0}{}^{A} k_A(\zeta,\bar\zeta)u^*T_A(\zeta,\bar\zeta)u/k_B(\zeta,\bar\zeta)u^*T_B(\zeta,\bar\zeta)u$$

$$(7.10)\qquad = \begin{cases}\{(1+\epsilon)/(1-\epsilon)\}^{n-1} & \text{for } u_1 \neq 0,\\ \{(1+\epsilon)/(1-\epsilon)\}^{n} & \text{for } u_1 = 0,\end{cases}$$

where ϵ *denotes an arbitrary constant number in the interval* (0, 1).

Proof. By the relative invariancies of k_D and T_D under $BH(D)$, it suffices to prove that (7.9) and (7.10) for R_ϵ and $R_{-\epsilon}$ in place of A and B are shown, respectively, since we have $d\zeta/dz \to E_n$ and $|J_\zeta(z)| \to 1$ for each mapping (7.4) or (7.5) via an A-approach. Therefore, (7.9) and (7.10) are obtained by Lemma 7.1.

Theorem 7.1. *Let* D *be a bounded SPCGR-NAH domain at* Q; *then we have, for any constant nonzero vector* $u \equiv (u_1, \ldots, u_n)^T$,

$$(7.11)\qquad \lim_{z\to Q}{}^{A} R_D(z; u) = -2/(n+1)$$

(*cf. Bergman* [3] *for* $n = 1$, *Fuks* [7] *for* $n = 2$),

$$(7.12)\qquad \lim_{z\to Q}{}^{A} C_D(z; u) = -1$$

(*cf. Fuks* [8] *for* $n = 2$) *and*

$$(7.13)\qquad \lim_{z\to Q}{}^{A} S_D(z) = -n.$$

Proof. Using Corollary 6.1, Lemma 7.2, (5.11) and (5.17), we conclude (7.11), (7.12) and (7.13), since $R_{-\epsilon}^{\alpha\beta}$, $R_{\epsilon}^{\alpha'\beta'}$, $R_{-\epsilon}$ and R_ϵ are biholomorphically equivalent to the unit hypersphere and ϵ can be taken as small as we need by taking sufficiently large numbers α, β, α' and β'. □

Now, we turn to compose another sort of domains of comparison, which is an immediate extension of domains of comparison due to Bergman for $n = 1$ [3, p. 38].

The set $U(r) \equiv \{z \mid |z_1 - r|^2 + \Sigma_{i=2}^n |z_i|^2 < r^2,\ r\text{: positive constant}\}$ and $B(r) \equiv \{z \mid |z_1 + r|^2 > r^2 + \Sigma_{i=2}^n |z_i|^2,\ r\text{: positive constant}\}$ are biholomorphically equivalent to the unit hypersphere E under the mappings

$$(7.14)\qquad z = \zeta/r - (1, 0, \ldots, 0)^T$$

and

$$(7.15)\qquad z = \zeta/(\zeta_1 + r) - (1, 0, \ldots, 0)^T,$$

whose Jacobian determinants tend to r^{-n} and $-r^{-n}$ for $\zeta \to 0$, respectively. $B(r)$ is similar to a Siegel domain of the second kind. If we consider the sections $U(r;\, t)$ and $B(R;\, t)$ restricted by the counter surface $\Sigma_{i=2}^{n} |\zeta_i|^2 = r^2 t$ $(0 \leq t < 1)$, we have

$$U(r;\, t) = \{\zeta_1 \,|\, |\zeta_1 - r| < r\sqrt{1-t}\} \subset B(R;\, t) = \{\zeta_1 \,|\, |\zeta_1 + R| > \sqrt{R^2 + r^2 t}\}$$

and thus $U(r) \subset B(R)$ and $\partial U(r) \cap \partial B(R) = \{0\}$ for $R \geq r$.

By the same procedure in the proof of Lemmas 7.1 and 7.2, we have the following Lemma 7.3 and Theorem 7.2.

Lemma 7.3. *If $R \geq r$, we have, for $n \geq 1$,*

$$\lim_{\zeta\to 0}{}^{A} \lambda^{(1)}_{B(R)}/\lambda^{(1)}_{U(r)} = \lim_{\zeta\to 0}{}^{A} \lambda^{(1)}_{U(R)}/\lambda^{(1)}_{U(r)} = (R/r)^{n-1} \tag{7.16}$$

and for any constant nonzero vector $u \equiv (u_1, \ldots, u_n)^T$

$$\begin{aligned} \lim_{\zeta\to 0}{}^{A} \lambda^{(2)}_{B(R)}(u)/\lambda^{(2)}_{U(r)} &= \lim_{\zeta\to 0}{}^{A} \lambda^{(2)}_{U(R)}(u)/\lambda^{(2)}_{U(r)}(u) \\ &= \begin{cases} (R/r)^{n-1} & \text{for } u_1 \neq 0, \\ (R/r)^{n} & \text{for } u_1 = 0. \end{cases} \end{aligned} \tag{7.17}$$

Theorem 7.2. *Let D be a bounded domain which has domains $U(r)$ and $B(r)$ $(U(r) \subset D \subset B(r))$ of comparison, then we have the same results as in Theorem* 7.1.

Example 7.1. (i) Let H be a Hartogs domain (complete multicircular domain with center at $(\psi(0), 0)^T$) $\{z \,|\, |z_1 - \psi(0)| < \psi(|z_2|),\ |z_2| < r,\ r > 0,\ \psi(\rho) \in C^2\text{-class and } \psi(0) > 0,\ \psi'(0) = 0,\ \psi''(0) < 0\}$. Set $\Psi(z, \bar z) \equiv |z_1 - \psi(0)|^2 - \psi^2(|z_2|)$ $(H \equiv \{z \,|\, \Psi < 0\})$. Then we have the Levi determinant $L(\Psi) = -\psi^2(0)\lambda'(0)$, where $\lambda(\rho^2) = \psi^2(\rho)$ and $\lambda'(0)$ denotes $d\lambda(x)/dx|_{x=0}$. Since $\lambda'(0) = \psi(0)\psi''(0) < 0$, then H is strictly pseudoconvex at 0. As H is expressed as

$$\{z \,|\, z_1 + \bar z_1 > (|z_1|^2 - \lambda'(0)|z_2|^2)/\psi(0) + o|z|^2\}$$

(about the origin), H is a SPCGR-NAH domain at 0. Therefore, Theorem 7.1 holds in this case, i.e., $\lim^A R_H(z;\, u) = -2/3$, $\lim^A C_H(z;\, u) = -1$ and $\lim^A S_H(z) = -2$.

(ii) Let us set $H' \equiv \{z \,|\, |z_1 - \psi(0)| < \psi(|z_2|),\ |z_2| < r,\ r > 0,\ \psi(\rho)$ $(0 \le \rho \le r)$ is a decreasing real valued continuous function which satisfies $\psi(0) - a + \sqrt{a^2 - \rho^2} \le \psi(\rho) \le \psi(0) + a - \sqrt{a^2 + \rho^2}$ and $\psi(0) > a > r\}$. Then H' has the domains of comparison: $U(a) = \{z \,|\, |z_1 - a|^2 + |z_2|^2 < a^2\}$ and $B(a) = \{z \,|\, |z_1 + a|^2 > a^2 + |z_2|^2\}$, since $\psi(0) - a + \sqrt{a^2 - \rho^2} \le \psi(\rho)$ and $\sqrt{a^2 + \rho^2} + \psi(\rho) \le \psi(0) + a$ imply $U(a) \subset H'$ and $H' \subset B(a)$, respectively, and $\partial U(a) \cap \partial B(a) \cap \partial H' = \{0\}$ is evident. Hence from Theorem 7.2 we have the same results as in (i).

Theorem 7.3. *If $A = U(r)$ and $B = U(R)$ (or $B = B(R)$) are domains of comparison such that $A \subset D \subset B$ and $\partial A \cap \partial B \cap \partial D = \{0\}$ for $r \le R$, then we have, for any nonzero vector $u \equiv (u_1, \dots, u_n)^T$,*

$$(7.18)\qquad 2\{1 - \nu(R/r)^{2(n-1)}\} \le \lim_{\zeta \to 0}{}^{A} R_D(\zeta; u) \le 2\{1 - \nu(r/R)^{2n}\},$$

$$(7.19)\qquad (n+1)\{1 - \nu(R/r)^{n-1}\} \le \lim_{\zeta \to 0}{}^{A} C_D(\zeta; u) \le (n+1)\{1 - \nu(r/R)^{n}\}$$

and

$$(7.20)\qquad n(n+1)\{1 - \nu(R/r)^{n-1}\} \le \lim_{\zeta \to 0}{}^{A} S_D(z) \le n(n+1)\{1 - \nu(r/R)^{n-1}\}.$$

Proof. From Lemma 7.3 and Corollary 6.1, we have the results.

8. On the Ricci scalar curvature.

Theorem 8.1. *In a bounded domain D we consider the quantity*

$$J_{D,p}(z, \bar{z}) \equiv J_{D,(p,1)}(z, \bar{z}) = \det(k_D^p(z, \bar{z}) \times T_D(z, \bar{z})) \quad (\textit{see } (4.22)).$$

(i) *For $p \ge (n+1)/n$, which is the case that the metric $ds^2_{D,p} \equiv ds^2_{D,(p,1)}$ can be defined (see Corollary 4.2), it holds that $\Delta \log J_{D,p}(z, \bar{z}) > 0$ for $z \in D$ and there is no fixed point $z^0 \in D$ such that $J_{D,p}(z, \bar{z}) \le J_{D,p}(z^0, \bar{z}^0)$ for $z \in D$, where Δ denotes the Laplace-Beltrami operator:* $\mathrm{Sp}\, T_D^{-1} D_z^* D_z$.

(ii) *If there exists a maximal point $z^0 \in D$ such that $J_{D,p}(z, \bar{z}) \le J_{D,p}(z^0, \bar{z}^0)$ for $z \in D$, then p must be smaller than $(n+1)/n$.*

Proof. Since $S_D(z) < n(n+1)$ holds for a bounded domain D,

$$\Delta \log J_{D,p} = \mathrm{Sp}\{T_D^{-1}(pnT_D - (R_{\bar{\alpha}\beta}))\} = pn^2 - S_D(z) > pn^2 - n(n+1) \ge 0$$

for $p \geq (n+1)/n$. If there exists a point $z^0 \in D$ such that $J_{D,p}(z, \bar{z}) \leq J_{D,p}(z^0, \bar{z}^0)$ for $z \in D$, then by the theorem of E. Hopf (see [22]) we obtain $J_{D,p} \equiv$ constant. Hence $pnT_D - (R_{\bar{\alpha}\beta}) = 0$ for $z \in D$ follows and thus $S_D(z) = pn^2 \geq n(n+1)$ for $p \geq (n+1)/n$, which is contradictory to (5.9). The proof of (ii) is clear.

Remark 8.1. [12, Theorem 3.10] says that in a bounded domain D, if there exists $z^0 \in D$ such that $J_D(z, \bar{z}) \leq J_D(z^0, \bar{z}^0)$ for $z \in D$, where $J_D \equiv k_D^{n+1} \det T_D$ $(= J_{D,(n+1)/n})$, then we have $J_D(z, \bar{z}) =$ constant and $S_D(z) = n(n+1)$. But this conclusion contradicts (5.9). Therefore, it seems to be faulty. This is also an impossible case of Theorem 8.1 (i) for $p = (n+1)/n$.

For $p = -1/n$, we have $J_{D,-1/n} = \det T_D/k_D \equiv I_D^{(1)}(z, \bar{z})$ which is a biholomorphically absolute invariant (see (4.13)). Thus the following theorem is an extension of [12, Theorem 3.9], which is obtained immediately by setting $p = -1/n$ in Theorem 8.1.

Theorem 8.2. *In a bounded domain D, let* $S_D(z) \geq s_0$ *(resp.* $S_D(z) \leq s_0$*) for* $z \in D$, *where* s_0 *is such a constant number that* $s_0 < n(n+1)$. *If for a real number* $p \leq s_0/n^2$ *(resp.* $p \geq s_0/n^2$*)* $J_{D,p}(z, \bar{z}) \geq J_{D,p}(z^0, \bar{z}^0)$ *(resp.* $J_{D,p}(z, \bar{z}) \leq J_{D,p}(z^0, \bar{z}^0)$*) in D holds for a fixed point* $z^0 \in D$, *then we have* $S_D(z) = s_0$ *in D.*

Proof. If $S_D(z) \geq s_0$ for $z \in D$ and $p \leq s_0/n^2$, we have $\Delta \log J_{D,p}(z, \bar{z}) = pn^2 - S_D(z) \leq pn^2 - s_0 \leq 0$ for $z \in D$. Therefore, if $J_{D,p}(z, \bar{z}) \geq J_{D,p}(z^0, \bar{z}^0)$ holds for $z \in D$, then from the theorem of E. Hopf we obtain $J_{D,p}(z, \bar{z}) =$ constant in D and thus $S_D(z) = pn^2 \leq s_0$. On the other hand, $S_D(z) \geq s_0$ holds from the hypothesis. Then we have $S_D(z) = s_0$ in D.

Theorem 8.3. *In a bounded homogeneous domain D, if* $J_{D,p}(z, \bar{z}) \geq J_{D,p}(z^0, \bar{z}^0)$ *(resp.* $J_{D,p}(z, \bar{z}) \leq J_{D,p}(z^0, \bar{z}^0)$*) holds in D, then we have* $J_{D,p}(z, \bar{z}) =$ *constant in D when and only when* $p = -1/n$, *i.e.,* $J_{D,p}(z, \bar{z}) \equiv I_D^{(1)}(z, \bar{z})$ *(see (4.13)).*

Proof. From (5.11) we have $S_D(z) = -n$. Therefore, if $J_{D,p}(z, \bar{z}) =$ constant, we have $\Delta \log J_{D,p}(z, \bar{z}) = pn^2 + n = 0$ and thus $p = -1/n$. On the other hand, if $p = -1/n$, we have $\Delta \log J_{D,p}(z, \bar{z}) = 0$ in D. Using the hypothesis and the theorem of Hopf, we obtain $J_{D,p}(z, \bar{z}) =$ constant.

Example 8.1. In the case of the first type $R(\mathrm{I})$ of the classical Cartan domains, which are homogeneous domains, we have

$$J_{R(\mathrm{I}),p}(z, \bar{z}) = k_{R(\mathrm{I})}^{pmn} \det T_{R(\mathrm{I})} = (m+n)^{mn}/V \det(E_m - z^* z)^{(m+n)(pmn+1)}$$

(dim $R(\mathrm{I}) = mn$), where V denotes the Euclidean volume of $R(\mathrm{I})$. Therefore, $J_{R(\mathrm{I}),p}(z, \bar{z}) = \text{constant} = (m+n)^{mn}/V$ holds when and only when $p = -1/mn$ (see [16]).

Theorem 8.4. *Let D be a bounded domain in C^2, whose Levi-expression $L(\phi)$ ($\phi \in C^2$-class) is positive at every point on D and let $I_D^{(1)}(z, \bar{z}) \equiv \det T_D(z, \bar{z})/k_D(z, \bar{z})$ be nonconstant. If there exists a point $z^0 \in D$ such that $I_D^{(1)}(z^0, \bar{z}^0) > 9\pi^2/2$ (resp. $I_D^{(1)}(z^0, \bar{z}^0) < 9\pi^2/2$), then $S_D(z)$ cannot be bounded by -2 from above (resp. below).*

Proof. In a bounded homogeneous domain G, $I_G^{(1)}(z, \bar{z}) = \text{constant}$ in G. Therefore, the domain D mentioned here is a nonhomogeneous domain. By the result of Bergman [3], $I_D^{(1)}(z, \bar{z})$ must assume its maximum (or minimum) in D with $L(\phi) > 0$. If there exists a point $z^0 \in D$ such that $I_D^{(1)}(z^0, \bar{z}^0) > 9\pi^2/2$, $I_D(z, \bar{z})$ must have its maximum in D. In this case, if $\Delta \log I_D^{(1)}(z, \bar{z}) = -2 - S_D(z) \geq 0$ in D, we have, by the theorem of Hopf, $I_D^{(1)}(z, \bar{z}) = \text{constant}$ in D. This is a contradiction. Therefore, $S_D(z)$ cannot be bounded by -2 from above.

9. **Reproducing kernel functions of subspaces.** Recently, B. L. Chalmers [5] has shown that the Riesz representation of any bounded linear functional in a Hilbert space with kernel function is obtained by operating with the linear functional on the kernel function itself and that, using this representation, one can display, in terms of the kernel function of the original space, the kernel function of any closed subspace defined as the intersection of the null spaces of at most countably many bounded linear functionals. In [5] he gives the following

Proposition 9.1. *Let $k_D(z, \bar{w})$ be the reproducing kernel function of a bounded domain D and $\mathfrak{L}_{(m)} \equiv (\mathfrak{L}_1, \ldots, \mathfrak{L}_m)$ be any bounded linear functionals with respect to z in D which are linearly independent. Then the kernel function of a subspace $\mathfrak{L}^2_{(m)}(D) = \{f \in \mathfrak{L}^2(D) | \mathfrak{L}_{(m)} f = K(m) \equiv (0, \ldots, 0)\}$ is given by*

$$k_{D,m}(z, \bar{w}) = \det \begin{pmatrix} k_D(z, \bar{w}), & \mathfrak{L}_{(m)} k_D(z, \bar{w}) \\ \mathfrak{L}^*_{(m)} k_D(z, \bar{w}), & \mathfrak{L}^*_{(m)} \mathfrak{L}_{(m)} k_D(z, \bar{w}) \end{pmatrix} \cdot (\det \mathfrak{L}^*_{(m)} \mathfrak{L}_{(m)} k_D(z, \bar{w}))^{-1}, \tag{9.1}$$

where $\mathfrak{L}^*_{(m)} \mathfrak{L}_{(m)} k_D(z, \bar{w}) = (\mathfrak{L}_{(m)} (\mathfrak{L}_{(m)} k_D(z, \bar{w}))^*)^*$ [5], [18].

The kernel function $k_D(z, \overline{w})$ has interesting minimalities as is well known (see (3.12)). We shall give another expression of $k_{D,m}(z, \overline{w}) \equiv k_m(z, \overline{w})$ as a minimizing function and show a sort of minimality of it by making use of the general minimum problem for $Q(z, \overline{z}) \equiv Q(z) = k_D(z, \overline{w})$.

Theorem 9.1. *For any fixed point $w \in D$, under the additional condition $Q(z, \overline{z}) \equiv Q(z) = k_D(z, \overline{w})$ and $\mathfrak{L}_{(m)}f = K(m) \equiv (0, \ldots, 0)$, we have the minimizing function*

$$(9.2) \qquad M^{(m)}_{D,k}(z, w) = k_D(z, \overline{w}) - \phi^*_D(w)\Phi_m(\Phi^*_m\Phi_m)^{-1}\Phi^*_m\phi_D(z) \in \mathfrak{L}^2_{(m)}(D),$$

where $\Phi_m = \mathfrak{L}_{(m)}\phi_D$ and $M^{(m)}_{D,k}(z, w) \equiv M^{K(m)}_{D,k_D(z,\overline{w})}(z, w)$ (3.3).

The function $M^{(m)}_{D,k}(z, w)$ coincides with the reproducing kernel function $k_m(z, \overline{w}) \in \mathfrak{L}^2_{(m)}(D)$ and equals the minimum value $\lambda^{K(m)}_{D,k_D(z,\overline{w})}(w)$ at (3.4) with $K(m) \equiv (0, \ldots, 0)$. Further $M^{(m)}_{D,k}(w, w) = k_m(w, \overline{w}) \leq k_D(w, \overline{w})$ holds.

Proof. In Theorem 3.1 if we set $Q(z, \overline{z}) \equiv k_D(z, \overline{w}) \in \mathfrak{L}^2(D)$ (w: fixed) and $\mathfrak{L}_{(m)}f = K(m) = (0)$, we have, from (3.3),

$$M^{(m)}_{D,k}(z, w) = \{B - B\Phi_m(\Phi^*_m\Phi_m)^{-1}\Phi^*_m\}\phi_D(z) \in \mathfrak{L}^2_{(m)}(D),$$

where $B = \int_D k_D(\zeta, \overline{w})\phi^*_D(\zeta)\omega_\zeta = \int_D \phi^*_D(w)\phi_D(\zeta)\phi^*_D(\zeta)\omega_\zeta \cdot = \phi^*_D(w)$. Noting $\phi^*_D(w)\phi_D(z) \equiv k_D(z, \overline{w})$, we have (9.2).

Since, for any $f(z) \in \mathfrak{L}^2_{(m)}(D)$,

$$\int_D f(\zeta)(\mathfrak{L}^*_{(m)}k_D(\zeta, \overline{w}))^*\omega_\zeta = \int_D f(\zeta)(\Phi^*_m\phi_D(\zeta))^*\omega_\zeta = \mathfrak{L}_{(m)}f = 0$$

follows from the Riesz's theorem, then we have

$$\int_D f(z)M^{(m)*}_{D,k}(z, w)\omega_z = \int_D f(z)k_D(w, \overline{z})\omega_z + 0 = f(w),$$

which shows that $M^{(m)}_{D,k}(z, w)$ has the reproducing property in $\mathfrak{L}^2_{(m)}(D)$. And further, $M^{(m)}_{D,k}(z, w)$ coincides with $k_m(z, \overline{w})$ by means of (9.1), since, in general, $\det\begin{pmatrix} a & B \\ C & D\end{pmatrix}(\det D)^{-1} = a - BD^{-1}C$ holds for a scalar a and a non-singular matrix D. Last parts of the theorem are easily obtained by (3.3) and (3.4).

Remark 9.1. If we set

$$\mathfrak{L}_{(m)} \equiv (\mathfrak{L}_{r(1),t_1}, \ldots, \mathfrak{L}_{r(m),t_m}),$$

where

$$\mathfrak{L}_{r(k),t_k} f \equiv d^{r(k)} f(z)/dz^{r(k)}\big|_{z=t_k} \qquad (k = 1, \ldots, m)$$

and

$$d^{r(k)}/dz^{r(k)} \equiv \partial^{r(k)}/\partial z_1^{r(k,1)} \cdots \partial z_n^{r(k,n)}$$

with $\Sigma_{i=1}^{n} r(k, i) = r(k) \geq 0$, we have another expression of Example 1.5 [5].

10. Fundamental theorem (I) of K. H. Look. In this section we shall give a neat but essentially equivalent proof of the fundamental theorem (I) given by K. H. Look [14] and an extension of this theorem using the minimum problem.

Proposition 10.1 (Fundamental theorem of Look). *Let D be a bounded schlicht domain and $f(z) \equiv (f_1(z), \ldots, f_n(z))^T$ be any holomorphic mapping with the condition $|f(z)| \leq M$ in D, then we have*

$$(10.1) \qquad (df(z)/dz)^*(df(z)/dz) \leq M^2 T_D(z, \bar{z}), \qquad z \in D,$$

and

$$(10.2) \qquad |J_f(z)|^2 \leq M^{2n} \det T_D(z, \bar{z}), \qquad z \in D.$$

Proof. Let $M_D^{K(2)}(z, t)$ be the minimizing function with the condition $Q(z, \bar{z}) \equiv 0$ and $K(2) = (A_1, A_2)$, and $F(z)$ be a holomorphic mapping $k_D(z, \bar{t}) f(z) \in \mathfrak{L}^2(D)$, then by (3.8), (3.9) and the Riesz's theorem for bounded linear functionals, we have

$$(10.3) \qquad \int_D F(z) \times M_D^{K(2)*}(z, t)\omega_z = f(t)A_1^* - f_1(t)T^{-1}(k_{10}A_1^* k^{-1} - A_2^*),$$

where $f_1(t) \equiv D_z f(t)$ and $T \equiv T_D(t, \bar{t})$. Setting $(A_1, A_2) = (0, T_D(t, \bar{t}))$ (this is possible since $f(z) = T_D(t, \bar{t})z$ belongs to $\mathfrak{L}^2_{(0,T)}(D)$), we have

$$f_1^* f_1 = T^* \left\{ \int_D M_D^{0E_n}(z, t) \times F^*(z)\omega_z \int_D F(z) \times M_D^{0E_n *}(z, t)\omega_z \right\} T.$$

For an arbitrary $n \times 1$ vector u, we have, by the Schwarz inequality,

$$u^* f_1^* f_1 u \leq \int_D |F(z)|^2 \omega_z \left(u^* T \int_D M_D^{0E_n} M_D^{0E_n *} \omega_z T u \right)$$

$$\leq k M^2 u^* T (kT)^{-1} T u = M^2 u^* T u.$$

This shows (10.1) and therefore (10.2).

Theorem 10.1. *Under the same hypothesis as in Proposition* 10.1, *we have*

$$(10.4)\qquad f(z)f^*(z) + f_1(z)T_D^{-1}(z, \bar{z})f_1^*(z) \leq M^2 \times E_n, \qquad z \in D,$$

and

$$(10.5)\qquad |J_f(z)|^2 \leq M^{2n} \det T_D(z, \bar{z})(1 - |f(z)|^2/M^2), \qquad z \in D.$$

If $f(z)$ *belongs to* $BH(D)$, *we have*

$$(10.6)\qquad f_1^*(z)\{E_n + f(z)f^*(z)/(M^2 - |f(z)|^2)\}f_1(z) \leq M^2T_D(z, \bar{z}), \qquad z \in D.$$

Proof. In (10.3), setting $F(z) = k_D(z, \bar{t})f(z)$, which belongs to $\mathfrak{L}^2(D)$, and $(A_1, A_2) = (F(t), dF(t)/dz)$, we have

$$\int_D F(z) \times M_D^{A_1A_2*}(z, t)\omega_z = k(ff^* + f_1T^{-1}f_1^*),$$

where $k \equiv k_D(t, \bar{t})$. By a way similar to that of the proof of Proposition 10.1, we obtain

$$k^2(ff^* + f_1T^{-1}f_1^*)^2 \leq M^2k^2(ff^* + f_1T^{-1}f_1^*).$$

By the diagonalization of Hermitian matrices, we have (10.4) and thus (10.5) (cf. (10.1) and (10.2)).

Let us assume that $f(z)$ belongs to $BH(D)$ in (10.4). Since $f_1T^{-1}f_1^* \leq M^2 \times (E_n - ff^*/M^2)$ follows from (10.4), we obtain $f_1^*(E_n - ff^*/M^2)^{-1}f_1 \leq M^2T$ by taking the inverse on both sides of the above. If A and B are positive definite Hermitian matrices and satisfy $A \leq B$, we have $A^{-1} \geq B^{-1}$, because, from a known theorem of matrices, A and B are simultaneously brought to diagonal matrices with positive diagonal elements by operating suitable regular matrices P^* and P on each of A and B as P^*AP and P^*BP. Noting that $(E_n - ff^*/M^2)^{-1} = E_n + ff^*/(M^2 - |f|^2)$, we get (10.6), which is an extension of (10.1) for $f(z) \in BH(D)$.

REFERENCES

1. S. Bergman, *Über die kernfunktion eines Bereiches und ihr Verhalten am Rande*, J. Reine Angew. Math. 169 (1933), 1–42; ibid. 172 (1934), 89–128.

2. ———, *Sur la fonction-noyau d'un domaine et ses application dans la théorie des transformations pseudoconformes*, Mém. Sci. Math., no. 108, Gauthier-Villars, Paris, 1948. MR 11, 344.

3. ———, *The kernel function and conformal mapping*, 2nd ed., Math. Surveys, no. 5, Amer. Math. Soc., Providence, R. I., 1970.

4. B. L. Chalmers, *On boundary behavior of the Bergman kernel function and related domain functionals*, Pacific J. Math. 29 (1969), 243–250. MR 40 #402.

5. ———, *Subspace kernels and minimum problems in Hilbert spaces with kernel function*, Pacific J. Math. 31 (1969), 619–628.

6. B. A. Fuks, *Über geodasische Mannigfaltigkeiten einer invarianten Geometrie*, Mat. Sb. 2(44) (1937), 567–594.

7. ———, *Special chapters in the theory of analytic functions of several complex variables*, Fizmatgiz, Moscow, 1963; English transl., Transl. Math. Monographs, vol. 14, Amer. Math. Soc., Providence, R. I., 1965. MR 30 #4979; 32 #5915.

8. ———, *Ricci curvature of a Bergman metric invariant under biholomorphic mappings*, Dokl. Akad. Nauk SSSR 167 (1966), 996–999 = Soviet Math. Dokl. 7 (1966), 525–529. MR 33 #4954.

9. L. Hörmander, *L^2 estimates and existence theorems for the $\bar\partial$ operator*, Acta Math. 113 (1965), 89–152. MR 31 #3691.

10. L. K. Hua, *On the estimation of the unitary curvature of the space of several complex variables*, Sci. Sinica 4 (1955), 1–26.

11. S. Kato, *Canonical domains in several complex variables*, Pacific J. Math. 21 (1967), 279–291. MR 35 #5659.

12. K. Kikuchi, *Canonical domains and their geometry in C^n*, Pacific J. Math. 38 (1971), 681–696. MR 46 #3836.

13. S. Kobayashi, *Geometry of bounded domains*, Trans. Amer. Math. Soc. 92 (1959), 267–290. MR 22 #3017.

14. K. H. Look, *Schwarz lemma and analytic invariants*, Sci. Sinica 7 (1958), 453–504. MR 21 #5028.

15. W. T. Martin, *On a minimum problem in the theory of analytic functions of several variables*, Trans. Amer. Math. Soc. 48 (1940), 351–357. MR 2, 86.

16. S. Matsuura, *On the normal domains and the geodesics in the bounded symmetric spaces and the projective space*, Sci. Rep. Gunma Univ. 15 (1966), 1–21.

17. ———, *Bergman kernel functions and the three types of canonical domains*, Pacific J. Math. 33 (1970), 363–384. MR 43 #560.

18. H. S. Shapiro, *Reproducing kernels and Beurling's theorem*, Trans. Amer. Math. Soc. 110 (1964), 448–458. MR 28 #2225.

19. J. M. Stark, *Minimum problems in the theory of pseudoconformal transformations and their application to estimation of the curvature of the invariant metric*, Pacific J. Math. 10 (1960), 1021–1038. MR 22 #12237.

20. T. Tsuboi and S. Matsuura, *Some canonical domains in C^n and moment of inertia theorems*, Duke Math. J. 36 (1969), 517–536. MR 41 #509.

21. W. Wirtinger, *Über eine Minimalaufgabe im Gebiete der analytischen Funktionen von mehreren Veränderlichen*, Monatsh. Math. Phys. 47 (1939), 426–431. MR 1, 10.

22. K. Yano and S. Bochner, *Curvature and Betti numbers*, Ann. of Math. Studies, no. 32, Princeton Univ. Press, Princeton, N. J., 1953. MR 15, 989.

DEPARTMENT OF MATHEMATICS, NAGOYA INSTITUTE OF TECHNOLOGY, NAGOYA, JAPAN

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

UNIQUENESS AND α-CAPACITY ON THE GROUP 2^{ω} (1)

BY

WILLIAM R. WADE

ABSTRACT. We introduce a class of Walsh series $\mathfrak{T}_\alpha^+$ for each $0 < \alpha < 1$ and show that a necessary and sufficient condition that a closed set $E \subseteq 2^\omega$ be a set of uniqueness for $\mathfrak{T}_\alpha^+$ is that the α-capacity of E be zero.

1. Introduction. A Walsh series $S \equiv \sum_{k=0}^\infty a_k\psi_k$ is said to *belong to the class* $\mathfrak{A}$ if

$$\lim_{n\to\infty} 2^{-n}S_{2^n}(x) = 0 \quad \text{for all } x \in 2^\omega; \tag{1}$$

where

$$S_N(x) \equiv \sum_{k=0}^{N-1} a_k\psi_k(x)$$

for $N = 0, 1, \ldots$.

Let $0 \le \alpha \le 1$ and for each positive integer k set $[k] = 2^n$ where n is the nonnegative integer determined by $2^n \le k < 2^{n+1}$. A Walsh series $S \equiv \sum_{k=0}^\infty a_k\psi_k$ is said to *belong to the class* $\mathfrak{T}_\alpha$ if

$$\sum_{k=1}^\infty a_k^2[k]^{\alpha-1} < \infty. \tag{2}$$

The Walsh series S is said to *belong to the class* $\mathfrak{T}_\alpha^+$ if in addition to (2) there exist integers $0 < n_1 < n_2 < \cdots$ such that

$$S_{2^{n_j}}(x) \ge 0 \quad \text{a.e.} \tag{3}$$

If $0 < \alpha < 1$ then $\mathfrak{T}_\alpha \subseteq \mathfrak{A}$, since by Schwarz's inequality

$$[2^{-n}S_{2^n}(x)]^2 \le 2^{-n\alpha} \sum_{k=0}^{2^n-1} a_k^2[k]^{\alpha-1}.$$

Received by the editors February 5, 1974 and, in revised form, May 2, 1974.

AMS (MOS) subject classifications (1970). Primary 42A56, 42A48; Secondary 42A62.

Key words and phrases. Haar functions, Walsh functions, sets of uniqueness, α-capacity.

(1) This research was partially sponsored by a University of Tennessee Faculty Research Grant.

Let $\mathfrak{B}$ be a certain class of Walsh series. A subset E of the group 2^ω is said to be a *set of uniqueness for* $\mathfrak{B}$ if $S \in \mathfrak{B}$ and $\lim_{n\to\infty} S_{2^n}(x) = 0$ for $x \in 2^\omega \sim E$ imply that S is the zero series.

For each Borel set $E \subseteq 2^\omega$ let $\mathfrak{M}(E)$ denote the set of all nonnegative Borel measures concentrated on E with total variation 1. Let $0 < \alpha < 1$. We associate with each measure $\mu \in \mathfrak{M}(E)$ a *potential function*

(4)
$$\bar{U}_\mu(x) = \int_{2^\omega} K_\alpha(x-y)\, d\mu(y)$$

where K_α is the nonnegative, lower semicontinuous, integrable function $\{x\}^{-\alpha}$ introduced in [6]. Let

$$W_\alpha(E) \equiv \inf\{W_\alpha^\mu(E) \colon \mu \in \mathfrak{M}(E)\},$$

where for each $\mu \in \mathfrak{M}(E)$,

(5)
$$W_\alpha^\mu(E) = \|\bar{U}_\mu\|_\infty.$$

Then E is said to be of *α-capacity zero* if $W_\alpha(E) = +\infty$.

Crittenden and Shapiro [3] have shown that a Borel set $E \subseteq 2^\omega$ is a set of uniqueness for $\mathfrak{A}$ if and only if E is countable. For each $\alpha \in (0, 1)$ we shall show that a closed set $E \subseteq 2^\omega$ is a set of uniqueness for $\mathfrak{T}_\alpha^+$ if and only if the α-capacity of E is zero. For a large class of null Walsh series which is contained in $\mathfrak{T}_\alpha^+$ see [8].

This author is indebted to Professor Victor L. Shapiro who first posed this problem in 1964 with $\mathfrak{T}_\alpha$ in place of $\mathfrak{T}_\alpha^+$. The analysis presented here would also solve the original problem if a group 2^ω analogue of Frostman's maximal principle were known. For this connection and a theorem concerning the trigonometric analogue of this problem see [1].

2. **Fundamental lemmas.** We begin this section quoting two results which are straightforward modifications of Theorem 2.9 and Lemma 3.2 in [6].

Lemma 1. *Let $E_1, E_2, \ldots$ be a nested sequence of closed subsets of 2^ω such that $E \equiv \bigcap_{n=1}^\infty E_n$ is a set of α-capacity zero. Then* $\lim_{n\to\infty} W_\alpha(E_n) = +\infty$.

Lemma 2. *Given a set $E \subseteq 2^\omega$ of positive α-capacity there is a measure $\mu \in \mathfrak{M}(E)$ such that its potential function is in $L^\infty(2^\omega)$ and satisfies*

(6)
$$\bar{U}_\mu(x) \geq W_\alpha(E)$$

for almost every $x \in E$.

The first lemma we prove is

Lemma 3. *If* $S = \sum_{k=0}^{\infty} a_k \psi_k \in \mathfrak{A}$ *and* c, d *are dyadic rationals in* $[0, 1]$, *then there is a Walsh series* $T \in \mathfrak{A}$ *and an integer* N *such that* $n > N$ *implies*

$$T_{2^n}(x) = S_{2^n}(x) \quad \textit{for } x \in [c, d) \tag{7}$$

and

$$T_{2^n}(x) \equiv 0 \quad \textit{for } x \notin [c, d).$$

To establish this result we define $j \circ k$ for each pair j, k of nonnegative integers by $\psi_{j \circ k} \equiv \psi_j \psi_k$. Let $P(x) = \sum_{k=0}^{M} \beta_k \psi_k(x)$ be the Walsh polynomial which is equal to 1 for $x \in [c, d)$ and equal to 0 elsewhere. Let $T = \sum_{k=0}^{\infty} \gamma_k \psi_k$ where

$$\gamma_k = \sum_{j=0}^{M} \beta_j a_{k \circ j}. \tag{8}$$

Then by Šneĭder [11, p. 285],

$$T_{2^n}(x) \equiv P(x) S_{2^n}(x), \qquad x \in 2^{\omega}, \tag{9}$$

when $2^n > M$. In particular, the choice of P forces T to have the desired properties.

Fine [4] has shown that a Walsh series S which converges to zero on an interval I with dyadic rational endpoints necessarily converges uniformly on I. It turns out that the 2^nth partial sums of S eventually *vanish* on I. In fact:

Lemma 4. *Let* F *be a closed subset of* $[0, 1]$ *and* $S = \sum_{k=0}^{\infty} a_k \psi_k \in \mathfrak{A}$. *Suppose further that* $\lim_{n \to \infty} S_{2^n}(x) = 0$ *a.e.* $x \in [0, 1] \sim F$ *and that* $\limsup_{n \to \infty} |S_{2^n}(x)| < \infty$ *for all but countably many* $x \in [0, 1] \sim F$. *Then for any interval* $(c, d) \subseteq [0, 1] \sim F$ *with dyadic rational endpoints there is an integer* N *such that* $n \geq N$ *and* $x \in (c, d)$ *imply* $S_{2^n}(x) \equiv 0$.

To prove Lemma 4 let T be the Walsh series corresponding to S and (c, d) given by Lemma 3. The conclusion of Lemma 3 and the hypotheses of Lemma 4 show us that T is a Walsh series, belonging to $\mathfrak{A}$, whose 2^nth partial sums converge to zero almost everywhere, are pointwise bounded off a countable set and satisfy (7) for n greater than some integer N. T is necessarily the zero series by the main theorem in [12]. Hence $S_{2^n}(x) \equiv 0$ for $x \in (c, d)$ and $n > M$ by (7).

3. The characterization.

Theorem. *Let $\alpha \in (0, 1)$ and E be a closed subset of the group 2^ω. Then a necessary and sufficient condition that E be a set of uniqueness for $\mathcal{T}_\alpha^+$ is that the α-capacity of E be zero.*

Necessity. Suppose the α-capacity of E is not zero. Then by definition there is at least one measure $\mu \in \mathfrak{M}(E)$ such that $W_\alpha^\mu(E) < \infty$. Let $d_0, d_1, \ldots$ represent the Walsh-Fourier-Stieltjes coefficients of μ and set $S = \Sigma_{k=0}^\infty d_k \psi_k$. S is not the zero series since $d_0 = \|\mu\| = 1$. Also, $\lim_{n\to\infty} S_{2^n}(x) = 0$ for $x \in [0, 1] \sim E$ since μ is supported on E [3, p. 563]. Furthermore $S_{2^n} \geq 0$ since $S_{2^n} = D_{2^n} * \mu$ and $D_{2^n} \geq 0$. Hence it suffices to show

Lemma 5. *Let $0 < \alpha < 1$, E be a closed subset of the group 2^ω, and $\mu \in \mathfrak{M}(E)$. Then there is a positive constant B depending only on α such that*

$$\sum_{k=0}^{\infty} d_k^2 [k]^{\alpha - 1} \leq W_\alpha^\mu(E) \cdot B \tag{10}$$

where $d_0, d_1, \ldots$ are the Walsh-Fourier-Stieltjes coefficients of μ.

Let $b_0, b_1, \ldots$ represent the Walsh-Fourier coefficients of K_α. Harper [6] has shown that there is a positive constant B depending only on α such that

$$Bb_k = [k]^{\alpha - 1}, \qquad k = 1, 2, \ldots. \tag{11}$$

For convenience let us define $[0]^{\alpha-1}$ so that (11) holds with $k = 0$.

To prove (10) we may suppose that $W_\alpha^\mu(E) < \infty$. In this case it is known that $\Sigma_{k=0}^\infty d_k^2 b_k = \int_{2^\omega} \bar{U}_\mu(x)\, d_\mu(x)$. Combining this with (11) we have

$$B^{-1} \sum_{k=0}^{\infty} d_k^2 [k]^{\alpha - 1} = \int_{2^\omega} \bar{U}_\mu(x)\, d_\mu(x) \leq W_\alpha^\mu(E).$$

Sufficiency. Suppose E is of α-capacity zero. Let $S = \Sigma_{k=0}^\infty a_k \psi_k$ be a Walsh series belonging to $\mathcal{T}_\alpha^+$ such that $\lim_{n\to\infty} S_{2^n}(x) = 0$ for $x \in 2^\omega \sim E$. We must show $a_k = 0$ for $k = 0, 1, \ldots$.

We first show that $a_0 = 0$. Let $\lambda\colon 2^\omega \to [0, 1]$ be defined by $\lambda(x_1, x_2, \ldots) = \Sigma_{k=1}^\infty x_k 2^{-k}$. Since λ is continuous [4] and E is compact, $\lambda(E)$ is necessarily closed in $[0, 1]$. Let $[0, 1] \sim \lambda(E) = \bigcup_{k=1}^\infty I_k$ where $I_1, I_2, \ldots$ is a sequence of open intervals with dyadic rational endpoints. Finally define a sequence of closed sets $E_1 \supset E_2 \supset \cdots$ in the group 2^ω by

$$E_N = 2^\omega \sim \lambda^{-1}\left(\bigcup_{k=1}^{N} I_k\right).$$

Now as a Walsh series on [0, 1], $\lim_{n\to\infty} S_{2^n}(x) = 0$ for $x \notin \lambda(E)$. Hence N applications of Lemma 4 allow us to conclude that for n sufficiently large, $S_{2^n}(x) \equiv 0$ for $x \in \bigcup_{k=1}^N I_k$. As a series on the group 2^ω, this means

$$S_{2^n}(x) \equiv 0 \quad \text{for } x \in 2^\omega \sim E_N. \tag{12}$$

Let $0 < n_1 < n_2 < \cdots$ satisfy (3). Since $a_0 \equiv \int_{2^\omega} S_{2^{n_j}}(x)\,dx$ we conclude that $a_0 \geq 0$. By (12),

$$|a_0| = \int_{E_N} S_{2^{n_j}}(x)\,dx \tag{13}$$

for j sufficiently large. Use Lemma 2 to choose an equilibrium measure $\mu \in \mathfrak{M}(E)$ satisfying (6). Then by (13) and (3)

$$|a_0| \leq \frac{1}{W_\alpha(E_N)} \int_{E_N} S_{2^{n_j}} \overline{U}_\mu(x)\,dx$$

for j sufficiently large. Hence by (12) and Parseval we have

$$|a_0| \leq \frac{1}{W_\alpha(E_N)} \sum_{k=0}^{2^{n_j}-1} a_k b_k d_k$$

for j sufficiently large, where $b_0, b_1, \ldots$ are the Walsh-Fourier coefficients of K_α and $d_0, d_1, \ldots$ are the Walsh-Fourier-Stieltjes coefficients of μ. Applying Schwarz's inequality we have

$$a_0^2 \leq W_\alpha^{-2}(E_N) \sum_{k=0}^{2^{n_j}-1} b_k^2 d_k^2 [k]^{(1-\alpha)} \cdot \sum_{k=0}^{2^{n_j}-1} a_k^2 [k]^{(\alpha-1)}.$$

But $S \in \mathfrak{T}_\alpha$ so by (11) and Lemma 5 we conclude that

$$a_0^2 \leq \text{const } W_\alpha^{-1}(E_N). \tag{14}$$

Observe that $\lambda^{-1} \circ \lambda(E) = \bigcap_{N=1}^\infty E_N$. Now $\lambda^{-1} \circ \lambda(E) \sim E$ is at most countable and the α-capacity of E is zero, so the α-capacity of $\bigcap_{N=1}^\infty E_N$ must also be zero. Hence by Lemma 1, $\lim_{N\to\infty} W_\alpha(E_N) = +\infty$, which by (14) implies that $a_0 = 0$.

For future reference, let us call what we have just proved a lemma.

Lemma 6. *Let $S \in \mathfrak{T}_\alpha^+$ and E be a closed set of α-capacity zero. If $\lim_{n\to\infty} S_{2^n}(x) = 0$ for $x \in 2^\omega \sim E$, then the constant term of S is zero.*

Suppose for some $m \geq 0$ that $a_k = 0$ for $k = 0, 1, \ldots, 2^m - 1$. We shall show that

$$a_k = 0 \quad \text{for } k = 2^m, 2^m + 1, \ldots, 2^{m+1} - 1 \tag{15}$$

thereby finishing the proof of the Theorem by induction.

To prove (15) fix an integer $l \in [2^m, 2^{m+1})$ and set

$$P(x) = D_{2^{m+1}}(l \cdot 2^{m+1} \dotplus x). \tag{16}$$

It is easy to see that $P(x) = \Sigma_{j=0}^{2^{m+1}-1} \beta_j^{(l)} \psi_j(x)$ where $\beta_j^{(l)} = \pm 1$ and that the matrix

$$A \equiv (\beta_j^{(l)}: j = 2^m, \ldots, 2^{m+1} - 1 \text{ and } l = 2^m, \ldots, 2^{m+1} - 1)$$

is nonsingular. For a similar result concerning Haar polynomials see [7].

Now set $T = \Sigma_{k=0}^{\infty} \gamma_k \psi_k$ where

$$\gamma_k = \sum_{j=0}^{2^{m+1}-1} \beta_j^{(l)} a_{k \circ j}. \tag{17}$$

A routine computation shows that $T \in \mathfrak{T}_a$. As in the proof of Lemma 3,

$$T_{2^n}(x) = P(x) S_{2^n}(x), \qquad x \in 2^\omega,$$

for n sufficiently large. In particular $\lim_{n \to \infty} T_{2^n}(x) = 0$ for $x \in 2^\omega \sim E$, and $T_{2^{n_j}}(x) \geq 0$ a.e., $j = 1, 2, \ldots$. Hence $\gamma_0 = 0$ by Lemma 6. By the inductive hypotheses and (17) we conclude $0 = \Sigma_{j=2^m}^{2^{m+1}-1} \beta_j^{(l)} a_j$. This identity holds for each $l = 2^m, 2^m + 1, \ldots, 2^{m+1} - 1$ so we finally arrive at the matrix equation

$$A \cdot \begin{bmatrix} a_{2^m} \\ \cdot \\ \cdot \\ \cdot \\ a_{2^{m+1}-1} \end{bmatrix} = 0.$$

Since the matrix A is nonsingular (15) is established as required.

BIBLIOGRAPHY

1. A. Broman, *On two classes of trigonometric series*, Ph. D. Dissertation, University of Upsala, 1947.

2. L. Carleson, *Selected problems on exceptional sets*, Van Nostrand Math. Studies, no. 13, Van Nostrand, Princeton, N. J., 1967. MR 37 #1576.

3. R. B. Crittenden and V. L. Shapiro, *Sets of uniqueness in the group* 2^{ω}, Ann. of Math. (2) **81** (1965), 550–564. MR 31 #3783.

4. N. J. Fine, *On the Walsh functions*, Trans. Amer. Math. Soc. **65** (1949), 372–414. MR **11**, 352.

5. B. Fuglede, *On the theory of potentials in locally compact spaces*, Acta Math. **103** (1960), 139–215. MR 22 #8232.

6. L. H. Harper, *Capacities of sets and harmonic analysis on the group* 2^{ω}, Trans. Amer. Math. Soc. **126** (1967), 303–315. MR 34 #6445.

7. J. R. McLaughlin and J. J. Price, *Comparison of Haar series with gaps with trigonometric series*, Pacific J. Math. **28** (1969), 623–627. MR 39 #3219.

8. F. Shipp, *Über Walsh-Fourier Reihen mit nicht-negativen Partialsummen*, Ann. Univ. Sci. Budapest Sect. Math. **12** (1969), 43–48.

9. V. L. Shapiro, *$U(\epsilon)$-sets for Walsh series*, Proc. Amer. Math. Soc. **16** (1965), 867–870. MR 31 #5035.

10. P. Sjölin, *An inequality of Paley and convergence a. e. of Walsh-Fourier series*, Ark. Mat. **7** (1969), 551–570. MR 39 #3222.

11. A. A. Šneĭder, *On the uniqueness of expansions for Walsh functions*, Mat. Sb. **24** (66) (1949), 279–300. (Russian) MR **11**, 352.

12. W. R. Wade, *A uniqueness theorem for Haar and Walsh series*, Trans. Amer. Math. Soc. **141** (1969), 187–194. MR 39 #4587.

13. ———, *Uniqueness of Haar series which are* $(C, 1)$ *summable to Denjoy integrable functions*, Trans. Amer. Math. Soc. **176** (1973), 489–498. MR 47 #704.

14. J. L. Walsh, *A closed set of normal orthogonal functions*, Amer. J. Math. **55** (1923), 5–14.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF TENNESSEE, KNOXVILLE, TENNESSEE 37916

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

POINTWISE BOUNDS ON EIGENFUNCTIONS AND WAVE PACKETS IN N-BODY QUANTUM SYSTEMS. III

BY

BARRY SIMON[1]

ABSTRACT. We provide a number of bounds of the form $|\psi| \le O(\exp(-a|x|^{\alpha}))$, $\alpha > 1$, for L^2-eigenfunctions ψ of $-\Delta + V$ with $V \to \infty$ rapidly as $|x| \to \infty$. Our strongest results assert that if $|V(x)| \ge cx^{2m}$ near infinity, then $|\psi(x)| \le D_\epsilon \exp(-(c-\epsilon)^{1/2}(m+1)^{-1}x^{m+1})$, and if $|V(x)| \le cx^{2m}$ near infinity, then for the ground state eigenfunction, Ω, $\Omega(x) \ge E_\epsilon \exp(-(c+\epsilon)^{1/2}(m+1)^{-1}x^{m+1})$.

1. Introduction. This is the last in our series of papers [19], [20] on pointwise bounds for L^2-eigenfunctions for Schrödinger operators $-\Delta + V$ on $L^2(\mathbf{R}^n)$. We have been partly motivated by a desire to extend and exploit the recent elegant techniques of O'Conner [15] and Combes-Thomas [3]. In (I) of this series, we considered the case $V = \sum V_{ij}(r_i - r_j)$ with $V_{ij}(x) \to 0$ as $x \to \infty$ and found exponential bounds $D_b \exp(-b|r|)$ but only for b smaller than some optimal b_0; in (II) of this series, we considered the case where V was bounded below and $V \to \infty$ as $r \to \infty$ and found exponential falloff for every b. In this paper, we wish to examine the case where V not only goes to infinity as $r \to \infty$ but at least as fast as some power r^{2m}. Not surprisingly, we will find that there is then falloff of $O(\exp(-cr^{\alpha}))$ for some $\alpha > 1$.

The relation between α and n is simple and is "predicted" by the following heuristic argument of WKB type [14]: If $\Delta\psi = W\psi$ and we write $\psi = \exp(-h)$, we find that h obeys

$$(\operatorname{grad} h)^2 - (\Delta h) = W.$$

If the variations of h are primarily radial we have $(\partial h/\partial r)^2 - r^{-2}(\partial/\partial r) \cdot (r^2(\partial h/\partial r)) = W$. If $W \to \infty$, then $\partial h/\partial r \to \infty$ so that the second derivate makes a small contribution. Thus $h \sim \pm \int W^{1/2}\, dr$, i.e.

Received by the editors May 17, 1974.

AMS (MOS) subject classifications (1970). Primary 81A45, 35B05; Secondary 31C05, 47A99.

(1) A. Sloan Foundation; research partially supported by The National Science Foundation under GP 39048.

$$\psi \sim \exp\left(-\int W^{1/2}\, dr\right).$$

If $W = r^{2m} - E$, we see that $\int W^{1/2} \sim r^{m+1}$, i.e. we expect to find that $\alpha = m + 1$.

For the case $n = 1$, it is often possible to use ordinary differential equation methods to control the falloff of eigenfunctions. For example, one has the following theorem of Hsieh-Sibuya [10] (see also the appendix by Dicke in [18]):

Theorem 1. *Let $\psi \in C^2(\mathbf{R})$ be a nonzero function obeying*

(1) $$-\psi'' + V\psi = E\psi$$

with

$$V(x) = a_{2m}x^{2m} + \cdots + a_0; \qquad a_{2m} > 0.$$

Then, for suitable c_0, either:

(a) $c_0\psi(x) \to \infty$ *as* $x \to +\infty$, *in which case* $(m+1)(\ln c_0\psi(x))/a_{2m}^{1/2}x^{m+1} \to 1$ *as* $x \to \infty$, *or*

(b) $c_0\psi(x) \to 0$ *as* $x \to +\infty$, *in which case* $(m+1)(\ln c_0\psi(x))/(-a_{2m}^{1/2}x^{m+1}) \to 1$ *as* $x \to \infty$.

The proof of Theorem 1 depends on the explicit construction of two independent solutions of (1) and thereby of all solutions. When $n > 1$, we have a partial differential equation and, in general, one cannot use a method listing all solutions. For later reference, we do note that in the case where V on $\mathbf{R}^n$ is centrally symmetric, one can separate variables in spherical coordinates and employ Theorem 1 to give some information.

We attack the problem of bounds on eigenfunctions of

(2) $$-\Delta\psi + V\psi = E\psi$$

by two methods. The first follows the approach of Combes-Thomas [3] and our earlier work [19], [20] and is discussed in §§2–4. We will be able to discuss fairly general V but our results wiii not always be as strong as might be hoped for. The second approach, found in §§5, 6 is completely independent of §§2–4 although it does depend on a result of Combes-Thomas type we proved in [20]. The V's we are able to discuss are somewhat restricted and so we restrict ourselves to multidimensional anharmonic oscillators, i.e. V will be a polynomial in $x_1, \ldots, x_n$ of degree $2m$ with the property that the leading term be strictly positive on the unit sphere (so that for x near ∞, $c_1|x|^{2m} \leq V(x) \leq c_2|x|^{2m}$). Our strongest result is (§6)

Theorem 2. *Let ψ be an L^2-eigenfunction for $-\Delta + V$. Suppose that V is C^∞ and for some $c > 0$, d:*

(3) $$V(x) \geq c|x|^{2m} - d.$$

Then for any $\epsilon > 0$ *there is a* D_ϵ *with*

$$|\psi(x)| \leq D_\epsilon \exp(-\sqrt{(c-\epsilon)}\,|x|^{m+1}(m+1)^{-1}). \tag{4}$$

Next suppose that ψ is the "ground state" eigenfunction, i.e. ψ is the eigenfunction associated to the lowest eigenvalue, E_0, of $-\Delta + V$. Then it is known (see, e.g. [22]) that E_0 is a nondegenerate eigenvalue and that ψ can be chosen to be a.e. strictly positive. For this ground state eigenfunction we have (§6)

Theorem 3. *Let* ψ *be the ground state eigenfunction for* $-\Delta + V$. *Suppose that* V *is* C^∞, $V \to \infty$ *at* ∞ *and for some* $e > 0$, f:

$$V(x) \leq e|x|^{2m} + f. \tag{5}$$

Then, for any $\epsilon > 0$, *there is a* G_ϵ *with*

$$\psi(x) \geq G_\epsilon \exp(-\sqrt{(e+\epsilon)}\,|x|^{m+1}(m+1)^{-1}). \tag{6}$$

In particular, ψ *is strictly positive.*

We close this introduction with a series of remarks about Theorems 2 and 3.

1. The proofs of Theorems 2 and 3 rely on Theorem 1 and a simple comparison argument (§5). The comparison argument depends on certain methods from classical potential theory; we have borrowed the idea of using these potential theory methods from Lieb-Simon [11] who is turn were motivated in part by some remarks of Teller [23].

2. Our interest in Theorem 3 and in the more general problem of sharp bounds on eigenfunctions of multidimensional anharmonic oscillators comes in part from recent work of Eckmann [5] and J. Rosen [17] generalizing L. Gross' logarithmic Sobolev inequalities [8]. We discuss the use of Theorem 3 to generalizing Rosen's results in §7.

3. Still another method for controlling falloff of eigenfunctions for anharmonic oscillators is to look at the finite dimensional Lie algebra generated by $-\Delta$ and V and use Lie algebraic techniques on eigenfunctions treated as analytic vectors. This approach has been advocated and developed by Goodman [6], [7] and Gunderson [9].

2. L^2 bounds of WKB type.

Theorem 4. *Let* $V = V_+ - V_-$ *with* $V_+ \geq 0$, $V_+ \in (L^1)_{\text{loc}}$, $V_- \in L^q(\mathbf{R}^n) + L^\infty(\mathbf{R}^n)$ *with* $q = 1$ *if* $n = 1$, $q > 1$ *if* $n = 2$ *and* $q = n/2$ *if* $n \geq 3$ *(so that* V_- *is a form bounded perturbation of* $-\Delta$ *with form bound* 0). *Let* $H = -\Delta + V$ *defined as a sum of quadratic forms. Let* ψ *be an eigenfunction for* H *with eigenvalue* E *in the discrete spectrum for* H. *Suppose* W *is a real-valued*

absolutely continuous function on $\mathbf{R}^n$ *with*

$$|\text{grad } W|^2 \le c_1(H + c_2) \tag{7}$$

for suitable c_1, c_2. *Then, for some* $\alpha > 0$,

$$\exp(\alpha W(x))\psi(x) \in L^2(\mathbf{R}^n). \tag{8}$$

Remarks. 1. In most applications, $V_+ \to \infty$ at ∞ so H has compact resolvent by Rellich's criterion. In such situations, E is automatically in the discrete spectrum.

2. As a particular example, suppose $V_- = 0$ and let $\tilde{V}(r) = \inf_{|x|=r} V(x)$. Then we can take $W(r) = \int_{r_0}^{r} |\tilde{V}(r)|^{1/2}\, dr$, thereby obtaining L^2-bounds on ψ of the usual WKB form.

3. Our proof is a fairly direct modification of the idea of Combes-Thomas [3] which in turn is motivated by [1], [2] (see also [21]).

Proof. For real β, let $U(\beta)$ be the unitary operator of multiplication by $\exp(i\beta W(x))$. (8) is easily seen to be equivalent to the statement that ψ be an analytic vector for $U(\beta)$ in the sense of Nelson. For β real, let

$$H(\beta) = U(\beta) H U(\beta)^{-1}.$$

Then

$$H(\beta) = (p - \beta \text{ grad } W)^2 + V \tag{9}$$

where $p = i^{-1}$ grad. Thus

$$H(\beta) = H + \beta^2(\text{grad } W)^2 - \beta[p(\text{grad } W) + (\text{grad } W)p]. \tag{9$'$}$$

Now, note the following estimates for $\phi \in Q(H) = Q(\Delta) \cap Q(V_+)$:

$$(\phi, (\text{grad } W)^2\phi) \le c_1(\phi, (H + c_2)\phi), \tag{10a}$$

$$2 \operatorname{Re}(p\phi, (\text{grad } W)\phi) \le 2(p\phi, p\phi)^{1/2}(\phi, (\text{grad } W)^2\phi)^{1/2} \le c_3(\phi, (H + c_4)\phi) \tag{10b}$$

where we have used (7) and the operator estimate $p^2 \le p^2 + (p^2 - 2V_- + c_5) \le 2(p^2 + V) + c_5$ which follows from the fact that V_- is a form perturbation of p^2 with relative bound 0.

Choose d with $H + d \ge 1$. It follows from (10(a)(b)) that for *complex* β sufficiently small, say $|\beta| \le B$, (9$'$) defines a closed sectorial form on $Q(H)$. It follows that for $|\beta| < B$, $H(\beta)$ is an analytic family of type (B) [12].

By analytic perturbation theory, it follows that for $|\beta| < B_0$, $H(\beta)$ has only discrete eigenvalues $E_1(\beta), \dots, E_n(\beta)$ in its spectrum near E and that the $E_i(\beta)$ are analytic. Since $H(\beta)$ is unitarily equivalent to H for β real, $E_i(\beta) = E$ for β real and thus, by analyticity for all β with $|\beta| < B_0$. Let

$$P(\beta) = \oint_{|\lambda - E| \leq \epsilon} (-2\pi i)^{-1}(H(\beta) - \lambda)^{-1}\, d\lambda$$

so that $P(\beta)$ is the projection onto the eigenvectors for $H(\beta)$ with eigenvalue E. Since $U(\alpha)P(\beta)U(\alpha)^{-1} = P(\beta + \alpha)$ for α real with $|\beta|, |\beta + \alpha| < B_0$, a lemma of O'Connor [15] assures us that $\psi \in \operatorname{Ran} P(0)$ is an analytic vector for $U(\alpha)$. □

3. Pointwise bounds, $m < 1$. We now wish to turn the L^2-bounds, $\psi \in D(\exp(\alpha W(x)))$, into pointwise bounds of the form

$$|\psi(x)| \leq A \exp(-\alpha' W(x)). \tag{11}$$

We consider the case $W(x) = |x|^{m+1}$. In this section, we will see how to use our method from [20] to obtain pointwise bounds in case $V_- = 0$ and $m \leq 1$. We note that our method in [20] was motivated by an idea of Davies [4]. We exploit smoothing properties of $\exp(t\Delta)$:

Lemma 3.1. *Let $\psi \in D(\exp(a|x|^{m+1}))$ for some $a > 0$ and $0 < m \leq 1$. Then for all t sufficiently small, there is an A and C (t dependent) so that*

$$|e^{t\Delta}\psi|(x) \leq C \exp(-A|x|^{m+1}).$$

Proof. We first note that

$$\begin{aligned} 1 + |x-y|^2 + |y|^{m+1} &\geq |x-y|^{m+1} + |y|^{m+1} \\ &\geq 2^{-m-1}(|x-y| + |y|)^{m+1} \geq 2^{-m-1}|x|^{m+1} \end{aligned}$$

so that

$$\exp(-a|x-y|^2)\exp(-a|y|^{m+1}) \leq \exp(1 - 2^{-m-1}a|x|^{m+1}).$$

Thus

$$\begin{aligned} &\left| \int \exp[-(a+1)|x-y|^2]\psi(y)\, dy \right| \\ &\qquad \leq \int \exp(-(a+1)|x-y|^2 - a|y|^{m+1})|\exp(a|y|^{m+1})\psi(y)|\, dy \\ &\qquad \leq \exp(1 - 2^{-m-1}a|x|^{m+1}) \int dy\, e^{-(x-y)^2}|\exp(a|y|^{m+1})\psi(y)|\, dy \\ &\qquad \leq C \exp(-2^{m-1}a|x|^{m+1}) \end{aligned}$$

since both factors in the integral are L^2. On account of the explicit form of the kernel for $e^{t\Delta}$, the lemma is proven. □

Theorem 5. *Let $V \in (L^2)_{\mathrm{loc}}$ with*

$$\alpha|x|^{2m} \leq V(x) + \beta \tag{12}$$

for suitable m, $0 < m \leq 1$, *and suitable* α, β. *Let* $H = -\Delta + V$ *defined as a selfadjoint operator sum* [13]. *Let* ψ *be an eigenfunction of* H. *Then, for some* $\gamma > 0$ *and* C:

$$|\psi(x)| \leq C \exp(-\gamma|x|^{m+1}). \tag{13}$$

Proof. By Rellich's criterion, (12) implies that H has only discrete spectrum. Letting $W(x) = |x|^{m+1}$ and using (12) and Theorem 4, we see that $\psi \in D(\exp(a|x|^{m+1}))$ for some $a > 0$.

Let V_k be a sequence of bounded functions with $V_k(x) - \beta$ converging montonically upward to V. Then using the fact that C_0^∞ is a common core [13], it is easy to see that $H_k \equiv -\Delta + V_k$ converges to H in strong resolvent sense [12], [16] as $k \to \infty$ so that $\exp(-tH)_k \to \exp(-tH)$ strongly as $k \to \infty$. Moreover, since $e^{t\Delta}$ is positivity preserving and $e^{t\beta} \geq e^{-tV_k} \geq 0$:

$$0 \leq (e^{-t\Delta/n}e^{-tV_k/n})^n|\phi| \leq e^{t\Delta}e^{t\beta}|\phi|$$

for all $\phi \in L^2$. By the Trotter product formula [16],

$$0 \leq e^{-tH_k}|\phi| \leq e^{t\beta}e^{t\Delta}|\phi|,$$

so by the convergence result:

$$0 \leq e^{-tH}|\phi| \leq e^{t\Delta}e^{t\beta}|\phi|.$$

Thus for any eigenfunctions ψ with $H\psi = E\psi$:

$$|\psi| = e^{tE}|e^{-tH}\psi| \leq e^{t(E+\beta)}e^{t\Delta}|\psi|.$$

By the lemma, and the fact noted above that $|\psi| \in D(\exp(a|x|^{m+1}))$ we obtain (13). □

4. **Pointwise bounds,** $m > 1$. When $m > 1$, we are not able to use the method of the last section to obtain pointwise bounds. Instead, we rely on Sobolev type estimates and therefore obtain results whose hypotheses depend on n, the dimension of space. We illustrate the ideas first in the special case $n \leq 3$ where only minimal additional hypotheses are needed.

Lemma 4.1. *Let* $f(x) = a(x^2 + 1)^{(m+1)/2}$ *on* $\mathbb{R}^n$. *If* $\psi \in L^2(\mathbb{R}^n)$ *and* ψ, $\Delta\psi \in D(e^f)$, *then for any multi-index* α *with* $\alpha \leq 2$, $D^\alpha\psi \in D(\exp[(1-\epsilon)f])$ *for all* $\epsilon > 0$. *In particular*, $\Delta(e^{(1-\epsilon)f}\psi) \in L^2$.

Proof. By a simple argument, we need only prove a priori estimates for $\psi \in C_0^\infty(\mathbb{R}^n)$. We note first that for any β:

$$\int e^{\beta f}|\nabla\psi|^2 = -\int \psi^*(\Delta\psi)e^{\beta f} - \int \psi^*(\beta e^{\beta f})\nabla f \cdot \nabla\psi. \tag{14}$$

Let $\beta < 1$, then since $e^{\beta f}\psi^*$, $\Delta\psi \in L^2$ and $\nabla f e^{\beta f}\psi^* \in L^2$, the R.H.S. of (14) is finite and thus $\nabla\psi \in D(e^{\beta f/2})$. We can now apply (14) when $\beta < 3/2$ to conclude the R.H.S. is finite so that $\nabla\psi \in D((3/4 - \epsilon)f)$. Repeating the argument, we see that $\nabla\psi \in D(\exp((1-\epsilon)f))$. From

$$\Delta(e^{\beta f}\psi) = e^{\beta f}\Delta\psi + 2\beta(\nabla f)e^{\beta f}\nabla\psi + [\Delta(e^{\beta f})]\psi$$

we conclude that $e^{\beta f}\psi \in D(\Delta)$ for $\beta < 1$ so that $D^\alpha(e^{\beta f}\psi) \in L^2$ if $|\alpha| \le 2$. Since $\nabla\psi \in D(e^{\beta f})$, we see that $D^\alpha\psi \in D(\exp((1-\epsilon)f))$. □

Theorem 6. *Suppose that the hypotheses of Theorem 4 hold with $n \le 3$ and $W(x) = |x|^{m+1}$. Suppose in addition that*

$$|V(x)| \le C_1 \exp(C_2|x|^\alpha) \tag{15}$$

with $\alpha < m + 1$. Then any eigenfunction ψ of $-\Delta + V$ obeys

$$|\psi(x)| \le C_3 \exp(-C_4|x|^{m+1}) \tag{16}$$

for suitable $C_3, C_4 > 0$.

Proof. By Theorem 4, $\psi \in D(\exp(af))$ for suitable $a > 0$ with $f = (1 + |x|^2)^{(m+1)/2}$. Since $\Delta\psi = V\psi - E\psi$, $\Delta\psi \in (\exp((a-\epsilon)f))$ on account of (15). Thus by Lemma 4.1, $e^{(a-\epsilon)f}\psi \in L^2(\mathbb{R}^n) \cap D(\Delta)$. By a Sobolev estimate, $e^{(a-\epsilon)f}\psi$ is a bounded continuous function, so (16) holds. □

For general n, we need

Lemma 4.2. *Let k be a positive integer and let $D^\alpha\psi, \Delta(D^\alpha\psi) \in D(\exp(f))$ with $f|x| = a(x^2 + 1)^{(m+1)/2}$ for $|\alpha| \le 2k$. Then $D^\alpha\psi \in D(\exp((1 - \epsilon)f))$ for all $\epsilon > 0$ and $|\alpha| \le 2(k + 1)$. In particular, $\Delta^{(k+1)}(e^{(1-\epsilon)f}\psi) \in L^2$.*

Proof. This follows immediately from Lemma 4.1. □

Theorem 7. *Fix n and m. Suppose the distributional derivatives $D^\alpha V$ for $|\alpha| \le 2[n/4 + 9/8]$ (where $[x] \equiv$ greatest integer less than or equal to x) are locally L^1 obeying*

$$|D^\alpha V| \le C_\alpha \exp(-D_\alpha|x|^{m+1-\epsilon}) \qquad (\epsilon > 0)$$

and that moreover

$$V(x) \ge C|x|^{2m} - D.$$

Then any eigenfunction ψ of $-\Delta + V$ obeys

$$|\psi(x)| \le A \exp(-B|x|^{m+1})$$

for suitable $A, B > 0$.

Proof. Similar to Theorem 6 but employing $-\Delta D^\alpha\psi + D^\alpha(V\psi) = ED^\alpha\psi$ as well as $-\Delta\psi + V\psi = E\psi$. □

5. **A comparison argument.** We now turn to a method of obtaining falloff information for eigenfunctions which is independent of and stronger than the results of §§2–4 but under stronger hypotheses. As we have already stated in the introduction, this method is motivated by [23], [11] although the basic idea is fairly standard. J. M. Combes (private communication) has informed me that T. Kato (unpublished) has used a not dissimilar idea in the one-dimensional case. The basic comparison theorem is

Theorem 8.(2) *Let S be a closed ball in $\mathbf{R}^n$. Suppose that f, g are functions C^∞ in a neighborhood of $\overline{\mathbf{R}^n\setminus S}$, and that*
(i) $\Delta|f| \le V|f|$ *all* $x \notin S$,
(ii) $\Delta|g| \ge W|g|$ *all* $x \notin S$,
(iii) $f, g \to 0$ *as* $x \to \infty$,
(iv) $W(x) \ge V(x) \ge 0$ *all* $x \notin S$,
(v) $|f(x)| \ge |g(x)|$ *all* $x \in \partial S$.
Then $|f(x)| \ge |g(x)|$ *all* $x \in S$.

Remark. (i), (ii) are intended in the sense of distributional inequalities.

Proof. Let $D = \{x \mid |f(x)| < |g(x)|\}$ and let $\psi = |g(x)| - |f(x)|$ on D, which is open. Then, on D,

$$\begin{aligned} \Delta\psi &\ge W|g| - V|f| && \text{(by (i), (iv))} \\ &\ge V(|g| - |f|) && \text{(by (iv))} \\ &\ge 0 && \text{(by } x \in D). \end{aligned}$$

Thus ψ is subharmonic on D and so takes its maximum value on $\partial D \cup \{\infty\}$. But $\psi \to 0$, at ∞ by (iii), at points $x \in \partial D \cap \partial S$ by (i) and at points $x \in \partial D\setminus\partial S$ by definition. Thus $\psi(x) \le 0$ on D. But, by definition, $\psi(x) > 0$ on D so D is empty. □

6. **Eigenfunctions of anharmonic oscillators.**

Lemma 6.1. *For any $m > 0$, $C > 0$, there exist an f and E so that* $-\Delta f + C(x^2+1)^m f = Ef$ *with*

$$0 < f(x) \le D_\epsilon \exp(-(C-\epsilon)^{+1/2}|x|^{m+1}/(m+1)^{-1}) \tag{17}$$

all x. Moreover, for suitable $D' > 0$

(2) Added in proof. H. Kalf has pointed out a similar result in P. Hartman and A. W. Winter, *Partial differential equations and a theorem of A. Kneser*, Rend. Cir. Mat. Palermo (2) 4 (1955), 237–255. MR 18, 214.

(17′) $$f(x) \geq D'_\epsilon \exp(-(C+\epsilon)^{1/2}(m+1)^{-1}(x)^{m+1}).$$

Proof. Let $H = -\Delta + (x^2+1)^m$. Choose f to be the ground state eigenfunction for H (which exists since H has purely discrete spectrum). Then f is a.e. nonnegative, so by the symmetry of H, f is spherically symmetric. Thus f obeys a suitable second order *ordinary* differential equation so that it is impossible that f and ∇f both vanish. But since $f \geq 0$ and C^∞ (elliptic regularity) $f = 0$ implies that $\nabla f = 0$ so f is strictly positive.

We claim that (17) holds near infinity and so everywhere. This follows either by appealing to a suitable generalization of Theorem 1 (since $|x|^{m-1/2}f$ obeys an equation similar to 1 but with an extra $C|x|^{-2}$ in the potential) or by appealing directly to Theorem 1, using Theorem 8 and an argument similar to that used in Theorem 2 below. □

We now repeat

Theorem 2. *Let V be a C^∞ function on $\mathbf{R}^n$ and let g be an eigenfunction of $-\Delta + V$. Suppose that $V(x) \geq c|x|^{2m} - d$ for some $c, d > 0$. Then, for any $\epsilon > 0$, there is a D_ϵ with*

(18) $$|g(x)| \leq D_\epsilon \exp(-(c-\epsilon)^{1/2}|x|^{m+1}/(m+1)^{-1}).$$

Remark. It is easy to replace C^∞ by C^p for suitable finite p.

Proof. Let $(-\Delta + V)g = Eg$. Given ϵ, find f with $[-\Delta + (c-\epsilon/2)|x|^{2m}]f = E_0 f$, $0 < f \leq D_\epsilon \exp(-(c-\epsilon)^{1/2}|x|^{m+1})$. Let $\widetilde{V} = (c-\epsilon/2)|x|^{2m} - E_0$; $\widetilde{W} = V - E$. Find a sphere S with $\widetilde{V} \geq \widetilde{W} \geq 0$ outside S. Since $f > 0$, f is bounded below on ∂S, so choose $\widetilde{f}$ a multiple of f with $|g| \leq \widetilde{f}$ on ∂S. By Kato's inequality [13]

$$\Delta|g| \geq \operatorname{Re}((\operatorname{sgn} g)\Delta g) = \operatorname{Re}(W|g|) = W|g|.$$

Finally, we note that by the exponential falloff inequalities on g [20], $g \to 0$ at ∞. Thus applying Theorem 8, $|g| \leq f$ outside S. (18) now follows. □

Now consider V which is C^∞ with $V \to \infty$ at ∞. By Rellich's criterion, $-\Delta + V$ has compact resolvent and so a lowest eigenvalue E_0. By a standard argument [22], E_0 is simple, and the corresponding eigenvector, ψ is a.e. positive. Following [23], [11] we first note

Lemma 6.2. *If ψ is a.e. positive, C^∞ with $-\Delta\psi = (-V+E)\psi$ with V C^∞, then ψ is everywhere strictly positive.*

Proof. Suppose that $\psi(0) = 0$. We will prove that ψ is identically zero near 0 violating the fact that ψ is a.e. positive. This will prove that $\psi(0) \neq 0$ and by similar argument that $\psi \neq 0$ for all x.

Thus, suppose $\psi(0) = 0$. Let $c(r) = \int_{|x|=r} \psi(x)\,d\Omega$. Then $c(r) \to 0$ as $r \to 0$ and

$$r^{n-1}\frac{dc}{dr} = \int_{|x|=r} \frac{\partial\psi}{\partial r}\,dS = \int_{|x|\le r} (\Delta\psi)\,d\tau$$

$$\le \max_{|x|\le r} (|V - E|) r^{n-1} \int_0^r c(x)\,dx.$$

Fix R_0 and let $D = \max_{|x|\le R_0}(|V - E|)$. Then for $0 < r \le R_0$

$$\frac{dc}{dr} \le D\int_0^r c(x)\,dx \le (Dr) \max_{0\le x\le r} c(x).$$

Since $c(0) = 0$: $c(r) \le (\tfrac{1}{2}Dr^2)\max_{0\le x\le r} c(x)$ so for $0 < r \le R$, $\max_{0\le x\le r} c(x) \le (\tfrac{1}{2}Dr^2)\max_{0\le x\le r} c(x)$.

Choosing r so small that $Dr^2 < 2$ and $0 < r < R$, we see that $\max_{0\le x\le r} c(x) = 0$ so that $\psi(x) = 0$ if $|x| < r$. □

We next repeat

Theorem 3. *Let ψ be the ground state eigenfunction for $-\Delta + V$ where V is C^∞ and $V \to \infty$. Suppose that $V(x) \le e|x|^{2m} + s$. Then for any $\epsilon > 0$, there is a G_ϵ with*

$$\psi(x) \ge G_\epsilon \exp(-\sqrt{(e+\epsilon)}\,|x|^{m+1}(m+1)^{-1}). \tag{19}$$

Proof. Let $f = \psi$, $\tilde{V} = V - E$ and let $W = (e + \epsilon/2)|x|^{2m} + s$. Let g be the ground state of $-\Delta + W$ with ground state energy $\tilde{E}$ and let $\tilde{W} = W - \tilde{E}$. Pick S so that $\tilde{W} \ge \tilde{V} \ge 0$ outside S. Since f is strictly positive and C^∞ by Lemma 6.2, choose a multiple $\tilde{g}$ of g with $f \ge \tilde{g}$ on ∂S. Then $f \ge \tilde{g}$ on $\mathbf{R}^n/S$ by Theorem 8. Thus, by Lemma 6.1, (19) follows. □

When V is a polynomial, we can say much more about the eigenfunctions.

Theorem 9. *Let V be a polynomial in n variables on $\mathbf{R}^n$ with $C(x^{2m} - 1) \le V(x) \le d(x^{2m} + 1)$ for $m \ge 1$. Let ψ be an L^2-eigenfunction for $-\Delta + V$. Then:*

(a) *ψ is a real-analytic function and has an analytic continuation to the entire space $\mathbf{C}^n$.*

(b) *For any $y \in \mathbf{R}^n$, $\epsilon > 0$,*

$$|\psi(x + iy)| \le C_{y,\epsilon} \exp[-(m+1)^{-1}(d-\epsilon)^{1/2}|x|^{m+1}]$$

for all $x \in \mathbf{R}^n$.

(c) *For any $\epsilon > 0$, there are constants E and F with*

$$|\psi(z)| \le E \exp(-F|z|^{m+1})$$

all $z \in \mathbf{C}^n$ *with* $\arg z_1 = \cdots = \arg z_n$ *and* $|\arg z_1| \le \pi/2(m+1) - \epsilon$.

(d) *For any* $\epsilon > 0$, *there are constants* G_1 *and* G_2 *with*

$$|\psi(z)| \le G_1 \exp(-G_2|z|^{m+1})$$

for all $z \in \mathbf{C}^n$ *with* $|\arg z_i| \le \pi/4m - \epsilon$, $i = 1, \ldots, n$.

Remark. With a minimal amount of extra work, one should be able to improve (d).

Proof. By the basic Combes-Thomas argument [3] we see that ψ is an entire analytic vector for the group $\{U(a)|a \in \mathbf{R}^n\}$ where $U(a)\psi(b) = \psi(b-a)$. Thus $\hat{\psi}$, the Fourier transform of ψ, has the property that $e^{ip\cdot a}\hat{\psi} \in L^2$ for all $a \in \mathbf{C}^n$. It follows that ψ is an entire function, proving (a). Moreover, $\psi(\cdot + iy)$ is an L^1-eigenfunction of $-\Delta + V(\cdot + iy)$ so the methods of §4 (or §5) allow one to prove (b). The bounds in (c), (d) follow by similar arguments (and a Phragmen-Lindelöf argument to get uniform constants) but using the group of dilations [1], [2], [21]. For (c) we note that $-\beta^{-2}\Delta + V(\beta x)$ is an analytic family of operators sectorial (in the sense of [16]) so long as $|\arg \beta| < \pi/2(m+1)$ and for (d) that

$$-\sum_{D=1}^{n} \beta_i^{-2} \frac{d^2}{dx_i} + V(\beta_1 x_1, \ldots, \beta_n x_n)$$

is accretive if $|\arg \beta_i| < \pi/4m$. □

Remark. Results related to Theorem 9 have been found by different methods in [9].

7. Supercontractive estimates à la J. Rosen. In [8], Gross considered the following situation. Let $H = -\Delta + V$ on $L^2(\mathbf{R}^n, dx)$ where V is a polynomial bounded from below. Let Ω be the ground state eigenfunction for H and let $\hat{H} = H - (\Omega, H\Omega)$. Let $d\mu$ be the probability measure $\Omega^2\,dx$. Then $\hat{H}$ on $L^2(\mathbf{R}^n, dx^n)$ is unitarily equivalent to $G = \Omega^{-1}\hat{H}\Omega$ on $L^2(\mathbf{R}^n, d\mu)$. G is a Dirichlet form in the sense that $(\psi, G\phi) = \int \overline{\operatorname{grad}\psi} \cdot \operatorname{grad}\phi\, d\mu$. Eckmann [5], following a suggestion of Gross [8], proved a variety of estimates which imply that G generates a hypercontractive semigroup [22] on $L^2(\mathbf{R}^n, d\mu)$ in case $n = 1$ or V is central and these estimates were improved by Rosen [17] who proved, in particular, that e^{-tG} is bounded from $L^p(\mathbf{R}^n, d\mu)$ to $L^q(\mathbf{R}^n, d\mu)$ for all $t > 0$, $p, q \neq 1, \infty$, again if $n = 1$. In Rosen's proof $n = 1$ enters in two places. First, he uses the fact that on $\mathbf{R}$, $f \le c(d^2/dx^2 + 1)$ if

$f \in L^1(\mathbf{R}, dx)$, but on $\mathbf{R}^n$ this can be replaced by $f \leq c(-\Delta + 1)$ if $f \in L^p(\mathbf{R}^n, dx)$, $p > n/2$ $(n > 2)$. More critically, he requires that $\Omega = e^{-h}$ with $(h)^{2m/m+1} \leq a(V + b)$ if $m = \frac{1}{2}\deg V$. This requires a lower bound on the falloff of Ω which was not available to him.

Our considerations in §6 were partially motivated by a desire to prove Rosen's estimates in case $n > 1$ and our results there allow us to mimic Rosen's proof [17] and conclude:

Theorem 10. *Let V be a polynomial on $\mathbf{R}^n$ with $a(x^{2m} - 1) \leq V(x) \leq b(x^{2m} + 1)$. Let $H = -\Delta + V$, Ω be its ground state, $d\mu = \Omega^2 d^n x$ and G be the Dirichlet form on $L^2(\mathbf{R}^n, d\mu)$. Then:*

(i) *For all $f \in C_0^\infty(\mathbf{R}^n)$:*

$$\int |f|^2 (\log_+ |f|)^{2km/m+1} d\mu \leq C_k \sum_{|\alpha| \leq k} \int |D^\alpha f|^2 d\mu + \|f\|_2^2 (\log \|f\|_2)^{2km/m+1}.$$

(ii) $D(G^{k/2}) \equiv \{f \in (C_0^\infty)' | D^\alpha f \in L^2;\ |\alpha| \leq k\}$.

(iii) *For all $t > 0$, p, $q \neq 1$, ∞, e^{-tG} is bounded from $L^p(\mathbf{R}^n, d\mu)$ to $L^q(\mathbf{R}^n, d\mu)$.*

Remark. By using the upper bounds we have on Ω, we can show that the inequality in (i) fails if a factor of $\log_q(\log_q(\cdots \log_q(|f|)))$ (j times) is added to the integral for any $j > 0$. This follows by Rosen's arguments [17].

REFERENCES

1. J. Aguilar and J. Combes, *A class of analytic perturbations for one-body Schrödinger Hamiltonians*, Comm. Math. Phys. 22 (1971), 269–279.

2. E. Balslev and J. Combes, *Spectral properties of many-body Schrödinger operators with dilatation-analytic interactions*, Comm. Math. Phys. 22 (1971), 280–294.

3. J. Combes and L. Thomas, *Asymptotic behavior of eigenfunctions for multiparticle Schrödinger operators*, Comm. Math. Phys. 34 (1973), 251–270.

4. E. B. Davies, *Properties of the Green's functions of operators*, J. London Math. Soc. (to appear).

5. J. P. Eckmann, *Hypercontractivity for anharmonic oscillators*, J. Functional Analysis (to appear).

6. R. W. Goodman, *Analytic and entire vectors for representations of Lie groups*, Trans. Amer. Math. Soc. 143 (1969), 55–76. MR 40 #1537.

7. ———, *Analytic domination by fractional powers of a positive operator*, J. Functional Analysis 3 (1969), 246–264. MR 39 #801.

8. L. Gross, *Logarithmic Sobolev inequalities*, Amer. J. Math. (to appear).

9. P. Gunderson, Thesis, Rutgers University, 1974 (in preparation).

10. P. Hsieh and Y. Sibuya, *On the asymptotic integration of second order linear ordinary differential equations with polynomial coefficients*, J. Math. Anal. Appl. 16 (1966), 84–103. MR 34 #403.

11. E. Lieb and B. Simon, *The Thomas-Fermi theory of atoms, molecules and solids* (in preparation).

12. T. Kato, *Perturbation theory for linear operators*, Die Grundlehren der math. Wissenschaften, Band 132, Springer-Verlag, New York, 1966. MR 34 #3324.

13. ———, *Schrödinger operators with singular potentials*, Israel J. Math. 13 (1972), 135–148.

14. V. P. Maslov, *Theory of perturbations and asymptotic methods*, Izdat. Moskov. Gos. Univ., Moscow, 1965; French transl., Dunod, Paris, 1972.

15. A. O'Conner, *Exponential decay of bound state wave functions*, Comm. Math. Phys. 32 (1973), 319–340.

16. M. Reed and B. Simon, *Methods of modern mathematical physics*. Vol. I, Functional Analysis, Academic Press, New York, 1972.

17. J. Rosen, *Logarithmic Sobolev inequalities and supercontractivity for anharmonic oscillators*, Thesis, Princeton University, 1974; paper (in preparation).

18. B. Simon, *Coupling constant analyticity for the anharmonic oscillator*, Ann. Phys. 58 (1970), 76–136.

19. ———, *Pointwise bounds on eigenfunction and wave packets in N-body quantum systems*, I, Proc. Amer. Math. Soc. 42 (1974), 395–401.

20. ———, *Pointwise bounds on eigenfunctions and wave packets in N-body quantum systems*, II, Proc. Amer. Math. Soc. 45 (1974), 454–456.

21. ———, *Quadratic form techniques and the Balslev-Combes theorem*, Comm. Math. Phys. 27 (1972), 1–9. MR 47 #9989.

22. B. Simon and R. Høegh-Krohn, *Hypercontractive semigroups and two dimensional self-coupled Bose fields*, J. Functional Analysis 9 (1972), 121–180. MR 45 #2528.

23. E. Teller, *On the stability of molecules in the Thomas-Fermi theory*, Rev. Mod. Phys. 34 (1962), 627–631.

DEPARTMENTS OF MATHEMATICS AND PHYSICS, PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08540

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

UNIQUENESS OF COMMUTING COMPACT APPROXIMATIONS

BY

RICHARD B. HOLMES, BRUCE E. SCRANTON AND JOSEPH D. WARD

ABSTRACT. Let H be an infinite dimensional complex Hilbert space, and let $\mathcal{B}(H)$ (resp. $\mathcal{C}(H)$) be the algebra of all bounded (resp. compact) linear operators on H. It is well known that every $T \in \mathcal{B}(H)$ has a best approximation from the subspace $\mathcal{C}(H)$. The purpose of this paper is to study the uniqueness problem concerning the best approximation of a bounded linear operator by compact operators. Our criterion for selecting a unique representative from the set of best approximants is that the representative should commute with T. In particular, many familiar operators are shown to have zero as a unique commuting best approximant.

Introduction. Let H be an infinite dimensional complex Hilbert space, and let $\mathcal{B}(H)$ (resp. $\mathcal{C}(H)$) be the algebra of all bounded (resp. compact) linear operators on H. It is well known [4], [6] that $\mathcal{C}(H)$ is proximinal in $\mathcal{B}(H)$, that is, for every $T \in \mathcal{B}(H)$ there exists a $C \in \mathcal{C}(H)$ such that $\|T - C\| = \operatorname{dist}(T, \mathcal{C}(H))$. It was shown, in [7], for arbitrary noncompact T that the set $\mathcal{P}(T)$ of best compact approximants to T has infinite dimension. From this proposition it can be deduced that c_0 viewed as a subspace of m has the same property. These spaces are the first "natural" proximinal subspaces known to the authors to have such a property. This phenomenon leads one to the question of finding a unique representative from $\mathcal{P}(T)$. Thus the purpose of this paper is to study the uniqueness problem concerning the best approximation of a bounded linear operator by compact operators. Our criterion for selecting a unique representative C_T from $\mathcal{P}(T)$ is that C_T should commute with T.

Now, in general, to satisfy our criterion for arbitrary T is not an easy task, since Lomonosov has shown [8] that any operator commuting with a nontrivial compact operator has a nontrivial invariant subspace. However, we recall from [7] that operators in the set $\mathcal{C}(H)^0 \equiv \{T \in \mathcal{B}(H) \mid \|T\| = \operatorname{dist}(T, \mathcal{C}(H))\}$ (anticompact operators) have, by definition, a commuting best compact approximant, namely 0. The anticompact operators have been considered by Coburn [2] and were termed "extremely noncompact." To study

Received by the editors August 17, 1973 and, in revised form, June 5, 1974.
AMS (MOS) subject classifications (1970). Primary 41A50, 41A65, 47B05; Secondary 47A30, 47B20, 47D20.

this situation in more detail, we introduce two classes of operators in $\mathcal{B}(H)$:

$ZUC = \{T \in \mathcal{B}(H) \mid 0$ is the unique compact operator that commutes with $T\}$

and

$ZUCA = \{T \in \mathcal{B}(H) \mid 0$ is the unique operator in $\mathcal{P}(T)$ that commutes with $T\}$.

Clearly, $ZUC \cap \mathcal{C}(H)^0 \subset ZUCA \subset \mathcal{C}(H)^0$ and, as we shall see, these inclusions are proper. The following fact, whose proof is omitted, constitutes the only general necessary condition known to us for membership in the classes ZUC or $ZUCA$.

Proposition 1. *An operator in $\mathcal{B}(H)$ cannot belong to ZUC or $ZUCA$ if it has a compact direct summand.*

In the first section of this paper we show that several classes of operators are in $ZUCA$ be virtue of being in $ZUC \cap \mathcal{C}(H)^0$. In the second section we provide criteria for a weighted shift to belong to the various operator classes $\mathcal{C}(H)^0$, ZUC, and $ZUCA$. In the final two sections we consider some counterexamples and open questions. Any terms not defined in this paper may be found in [5].

At this time we would like to thank Professor C. R. Putnam for his many helpful discussions.

1. Operators in $ZUC \cap \mathcal{C}(H)^0$. What sort of operators are in $ZUCA$? Many operators are in $ZUCA$ by virtue of being in $ZUC \cap \mathcal{C}(H)^0$. We begin the investigation of this latter subset by identifying a large class of operators in $\mathcal{C}(H)^0$.

Let $r_e(T)$ be the essential spectral radius of $T \in \mathcal{B}(H)$. Although there are several notions of essential spectrum, it was shown in [9] that the corresponding essential spectral radii are all the same. Hence $r_e(T)$ is unambiguously defined as, for example, $\max\{|\lambda| \mid \lambda \in \bigcap_{C \in \mathcal{C}(H)} \text{Spectrum}(T + C)\}$.

Definition. $T \in \mathcal{B}(H)$ is *essentially normaloid* if $r_e(T) = \|T\|$.

In [7] it was observed that seminormal operators with empty point spectrum are essentially normaloid, and the following proposition was proved:

Proposition 2. *Every essentially normaloid operator is anticompact.*

Our strategy for this section may now be described. We will use Proposition 1 to restrict our attention to certain essentially normaloid operators. Then, in view of Proposition 2, to prove that such an operator is in $ZUCA$ it suffices to show that the operator belongs to ZUC.

Theorem 1. *A normal operator is in $ZUC \cap \mathcal{C}(H)^0$ if and only if its point spectrum is empty.*

Proof. Since any eigenspace of a normal operator is a reducing subspace, a normal operator with an eigenvalue has a compact direct summand and by Proposition 1 is not in $ZUCA$.

Conversely, let N be a normal operator with empty point spectrum. By the preceding discussion it is sufficient to show that $N \in ZUC$. Suppose that C is a compact operator and C commutes with N (written $C \leftrightarrow N$). We show $C = 0$. Now $N \leftrightarrow C$ implies $N \leftrightarrow C^*$ (Fuglede's theorem). Thus $N \leftrightarrow C$ implies $N \leftrightarrow C^*C$. Since C^*C is a positive, compact operator, the Schmidt (polar) decomposition asserts that the spectrum of C^*C consists of 0 and a (possibly empty) decreasing sequence of positive eigenvalues, each of finite multiplicity.

Suppose that E is an eigenspace of C^*C corresponding to a positive eigenvalue. It is easy to check that $N \leftrightarrow C^*C$ implies E is an invariant subspace of N. Since E is finite dimensional, this means that N must have an eigenvalue, which contradicts our hypothesis. Thus the spectrum of C^*C is $\{0\}$. Hence $C^*C = 0$, which implies $C = 0$. Q.E.D.

Theorem 2. *An isometry is in $ZUC \cap \mathcal{C}(H)^0$ if and only if its point spectrum is empty.*

Proof. Express the isometry in its Wold decomposition [5] as $U \oplus W$, where U is a pure isometry (i.e. a unilateral shift of some multiplicity) and W is a unitary operator. Any eigenspace of the isometry must be an eigenspace of the unitary part, and hence a reducing subspace of the isometry. Thus if an isometry has an eigenvalue, it has a compact direct summand, and by Proposition 1 it is not in $ZUCA$.

Conversely, if the point spectrum of the isometry (a subnormal operator) is empty, Proposition 2 is applicable, and it is sufficient to show that the isometry is in ZUC.

First, consider a pure isometry U. U is defined by

$$U(x_1, x_2, x_3, \ldots) = (0, x_1, x_2, x_3, \ldots)$$

where the x_j are elements of a fixed Hilbert space K such that $\Sigma \|x_j\|^2 < \infty$. Let $x \in K$ be a fixed unit vector, and define $e_n = (0, \ldots, 0, x, 0, \cdots)$ where x is the nth component of e_n. Then $\{e_n\}_{n=1}^{\infty}$ is an orthonormal sequence in the domain of U. Suppose C is a compact operator and $C \leftrightarrow U$. Then

$$UC(e_n) = CU(e_n) = C(e_{n+1}),$$

which implies

$$\cdots = \|C(e_{n+1})\| = \|C(e_n)\| = \cdots = \| C(e_1)\|.$$

Because C is compact, $\lim C(e_n) = 0$, and hence $C(e_n) = 0$ for every n; that is, $C = 0$.

Consider any compact operator $\hat{C}$ which commutes with the isometry. Corresponding to the Wold decomposition $\begin{bmatrix} U & 0 \\ 0 & W \end{bmatrix}$ of the isometry we have $\hat{C} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, where A, B, C, and D are compact. From the commutativity of these operators it follows that $A \leftrightarrow U$ and $D \leftrightarrow W$, so by the above paragraph and Theorem 1 we have $A = 0$ and $D = 0$. Further, $CU = WC$, and if we consider e_n as above we have

$$WC(e_n) = CU(e_n) = C(e_{n+1})$$

and $\cdots = \|C(e_{n+1})\| = \|C(e_n)\| = \cdots = \|C(e_1)\|$. As before, the compactness of C implies that $C = 0$. Lastly, $BW = UB$, so that $W^*B^* = B^*U^*$. Again letting e_n be as above, and recalling that U^* is the backwards shift we have

$$W^*B^*(e_{n+1}) = B^*U^*(e_{n+1}) = B^*(e_n)$$

and

$$\cdots = \|B^*(e_{n+1})\| = \|B^*(e_n)\| = \cdots = \|B^*(e_1)\| = 0,$$

so that $B^* = 0$, whence $B = 0$. Q.E.D.

Before proceeding to the last classes of operators in $ZUC \cap \mathcal{C}(H)^0$, we state and prove a proposition that will be used to show that the operators are in ZUC. The fact that ZUC and $ZUCA$ are invariant under adjunction is easy to verify and is used in the proposition.

Proposition 3. *If an operator has empty point spectrum and its adjoint has so many simple eigenvalues that the corresponding eigenvectors are fundamental in H, then the adjoint of the operator (hence the operator itself) is in ZUC.*

Proof. Suppose $C \leftrightarrow T$ and C is compact. By an argument similar to the one used in the proof of Theorem 1, it is clear that $\text{spectrum}(C) = \{0\} = \text{spectrum}(C^*)$. We show that $C^* = 0$ by showing $C^*(x) = 0$ for any eigenvector x associated with a simple eigenvalue λ of T^*. Since $C \leftrightarrow T$, we have $C^* \leftrightarrow T^*$ so that $T^*C^*(x) = C^*T^*(x) = \lambda C^*(x)$. Since λ is a simple

eigenvalue of T^*, x must be an eigenvector of C^*. Because spectrum $(C^*) = \{0\}$, we have $C^*(x) = 0$. Q.E.D.

Theorem 3. *Each of the following (classes of) operators is contained in* $ZUC \cap \mathcal{C}(H)^0 \subset ZUCA$:

(a) *the discrete Cesaro operator,*

(b) *multiplication by a bounded schlicht function on some Bergman space,*

(c) *Toeplitz operators whose corresponding multiplication function is schlicht.*

Proof. It is well known that these operators are subnormal and have empty point spectrum; thus, in accord with the strategy of this section, it is sufficient to show that they belong to ZUC. This we will do by showing that in each of these cases the hypotheses of Proposition 3 are satisfied.

Proof of (a). In [1] the following facts were proved: the point spectrum of the adjoint of the discrete Cesaro operator is $\{\lambda \mid |1 - \lambda| < 1\}$; each of these eigenvalues is simple; when l^2 is identified with the Hardy space H^2 in the natural manner, the function $(1 - z)^{1/\lambda - 1}$ is an eigenvector associated with λ. It remains to show that these eigenvectors are fundamental. By considering $\lambda = 1, 1/2, 1/3, \cdots$ it is easy to see that the span of the eigenvectors includes $1, z, z^2, \cdots$. Thus the span of the eigenvectors of the adjoint of the discrete Cesaro operator is dense. Q.E.D.

Proof of (b). Let

D = a fixed region in the complex plane,

ϕ = a bounded schlicht function on D,

T = multiplication by ϕ on $A^2(D)$,

K_λ = reproducing element for "evaluation at λ" functional δ_λ.

Since $\{K_\lambda\}_{\lambda \in D}$ is fundamental in $A^2(D)$, it is sufficient to show that $\overline{\phi(\lambda)}$ is a simple eigenvalue of T^* with corresponding eigenvector K_λ, for each $\lambda \in D$. To do this recall that

$$\ker (T^* - \overline{\phi(\lambda)}I) = \operatorname{ran} (T - \phi(\lambda)I)^\perp.$$

Thus, using the definition of K_λ, it is easy to check that K_λ is an eigenvector associated with $\overline{\phi(\lambda)}$. To see that $\overline{\phi(\lambda)}$ is simple we verify that $\operatorname{ran} (T - \phi(\lambda)I)$ is the kernel of a linear functional, viz.,

$$\operatorname{ran} (T - \phi(\lambda)I) = \{g \in A^2(D) | g(\lambda) = 0\} = \ker \{\delta_\lambda\}.$$

Now we clearly have

$$\operatorname{ran}(T - \phi(\lambda)I) = \{g \mid g(z) = (\phi(z) - \phi(\lambda))f(z) \text{ for some } f \in A^2(D)\}$$
$$\subset \{g \in A^2(D) \mid g(\lambda) = 0\}.$$

For any $g \in A^2(D)$ such that $g(\lambda) = 0$ we may define $f(z) = g(z)/(\phi(z) - \phi(\lambda))$, and the problem reduces to showing $f \in A^2(D)$. f is defined at $z = \lambda$ since

$$\lim_{z \to \lambda} \frac{g(z)}{\phi(z) - \phi(\lambda)} = \frac{g'(\lambda)}{\phi'(\lambda)}$$

and $\phi'(\lambda) \neq 0$ because ϕ is schlicht [11, p. 198]. It is similarly easy to check that f is differentiable at $z = \lambda$. To see that $f \in L^2(D)$, note that f is continuous on a disc D_λ centered at λ and contained in D. Thus f is certainly in $L^2(D_\lambda)$. It suffices to show that $|\phi(z) - \phi(\lambda)|$ is bounded away from 0 on $D \setminus D_\lambda$. If this were not true, there would exist z_n, $n = 1, 2, \ldots$, in $D \setminus D_\lambda$ such that $\phi(z_n) \to \phi(\lambda)$ as $n \to \infty$. Since ϕ^{-1} is also analytic on D [11, p. 199], $z_n \to \lambda$ as $n \to \infty$. This is a contradiction. Q.E.D.

Proof of (c). Using the representation of the Hardy space as $H^2(D)$ where D is the open unit disc, the proof is essentially the same as in part (b). The only difference is that for $g \in H^2(D)$ such that $g(\lambda) = 0$, it must be observed that $\int_0^{2\pi} |f(re^{i\theta})|^2 \, d\theta$ is uniformly bounded for r sufficiently close to but less than 1, where $f(z) = g(z)/(\phi(z) - \phi(\lambda))$. The proof of this observation is also analogous to the corresponding one in part (b). Q.E.D.

Remark 1. The above classes of operators in ZUC ($ZUCA$) are all hyponormal (even subnormal) and have empty point spectrum. From Proposition 1 and the fact that eigenspaces reduce hyponormal operators it follows that the empty point spectrum assumption was necessary for such operators to be in ZUC ($ZUCA$). However, this necessary condition breaks down for seminormal operators. For example, the adjoint of the unilateral shift is in ZUC (and $ZUCA$) by Proposition 2; yet its point spectrum is the open unit disc.

Remark 2. Although the result of Shields and Wallen [10, Theorem 2] implies that their multiplication operators M_z belong to ZUC, Proposition 3 is applicable to a more general situation where their condition (c) is significantly weakened and condition (d) is eliminated. We also mention that Theorem 3(c) has recently been proved independently by Deddens and Wong [3].

2. Weighted shifts. In this section we consider the following question: Which weighted shifts belong to the classes $\mathcal{C}(H)^0$, ZUC, and $ZUCA$? We will use the following notation for a weighted shift throughout this section:

$$T = \sum_{n=1}^{\infty} \alpha_n e_{n+1} \otimes \overline{e}_n$$

i.e.

$$T(x) = \sum_{n=1}^{\infty} \alpha_n \langle x, e_n \rangle e_{n+1}$$

where $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis of H, chosen in such a way that the weights α_n are nonnegative. If T had a zero weight it would have a finite rank direct summand, and by Proposition 1 it would not be in ZUC or $ZUCA$. Hence, we will require that all the weights be positive.

We begin by characterizing the weighted shifts in ZUC. For $T \in \mathcal{B}(H)$ a necessary condition for T to be in ZUC is that T^n be noncompact for every positive integer n. It is interesting to note that for weighted shifts this condition is also sufficient.

Theorem 4. *A weighted shift T with positive weights α_n belongs to ZUC if and only if there does not exist a $k_0 > 1$ so that* $\lim_n (\alpha_{n+1} \cdots \alpha_{n+k_0-1}) = 0$.

Proof. If there exists $k_0 > 1$ such that $\lim_n (\alpha_{n+1} \cdots \alpha_{n+k_0-1}) = 0$, then by the Schmidt decomposition T^{k_0-1} is compact and T is not in ZUC.

Conversely, suppose $C \leftrightarrow T$. This is equivalent to

$$TC(e_n) = CT(e_n) = \alpha_n C(e_{n+1}) \quad \text{for all } n.$$

Hence

$$C(e_{n+1}) = \frac{T}{\alpha_n} C(e_n) = \cdots = \frac{T^n}{\alpha_n \cdots \alpha_1} C(e_1) \quad \text{for all } n.$$

If T is not in ZUC, then we may assume that the above C is compact and nonzero. Thus $C(e_1) \neq 0$ and we may write $C(e_1) = \sum_{j=k_0}^{\infty} \beta_j e_j$ with $\beta_{k_0} \neq 0$. Because $\|T^n(C(e_1))\| \geq |\beta_{k_0} \alpha_{k_0} \cdots \alpha_{n+k_0-1}|$, and C is compact we have

$$0 = \lim_n \|C(e_{n+1})\| = \lim_n \frac{1}{\alpha_n \cdots \alpha_1} \|T^n(C(e_1))\|$$

$$\geq \lim_n |\beta_{k_0}| \frac{\alpha_{k_0} \cdots \alpha_{n+k_0-1}}{\alpha_n \cdots \alpha_1} = \frac{|\beta_{k_0}|}{\alpha_{k_0-1} \cdots \alpha_1} \lim_n \alpha_{n+1} \cdots \alpha_{n+k_0-1}.$$

Hence from the term immediately after the inequality we see that $k_0 > 1$, and it follows that $\lim_n (\alpha_{n+1} \cdots \alpha_{n+k_0-1}) = 0$. Q.E.D.

The following remarks will be useful later, and refer to a weighted shift T with positive weights.

Remark 3. If $T \notin \mathcal{C}(H)$ and $T \notin ZUC$, then the k_0 in this theorem satisfies $k_0 \geq 3$.

Remark 4. If $0 \neq C \in \mathcal{C}(H)$ and $C \leftrightarrow T$, then $C(e_1) = \Sigma_{j=k_0}^{\infty} \beta_j e_j$ with $\beta_{k_0} \neq 0$ and $k_0 > 1$. Thus for $n \geq 1$,

$$C(e_{n+1}) = \frac{T^n}{\alpha_n \cdots \alpha_1} C(e_1) = \frac{1}{\alpha_n \cdots \alpha_1} \sum_{j=k_0}^{\infty} \beta_j (\alpha_j \cdots \alpha_{j+n-1}) e_{j+n},$$

so that $C(e_{n+1})$ is orthogonal to $e_1, \ldots, e_{n+k_0-1}$.

Remark 5. If $C \in \mathcal{C}(H)$ and $C \leftrightarrow T$, then $C = 0$ if and only if $C(e_n) = 0$, for some integer n.

In [7] a characterization of the weighted shifts with nonnegative weights in $\mathcal{C}(H)^0$ was given, namely:

Proposition 4. *A weighted shift with nonnegative weights* α_n *belongs to* $\mathcal{C}(H)^0$ *if and only if* $\sup_n \alpha_n = \limsup_{n\to\infty} \alpha_n$.

Thus combining Propositions 3 and 4, we obtain a characterization of all weighted shifts in $ZUC \cap \mathcal{C}(H)^0$ in terms of the weights. From this characterization it may be easily verified that

Corollary. *A hyponormal weighted shift is in* $ZUC \cap \mathcal{C}(H)^0$ *if and only if its point spectrum is empty.*

We do not know a necessary and sufficient condition for a weighted shift to be in $ZUCA$. We do know, however, that the weighted shifts in $ZUC \cap \mathcal{C}(H)^0$ do not exhaust the weighted shifts in $ZUCA$. The next proposition will enable us to exhibit such an example.

Proposition 5. *Let* $T \in \mathcal{C}(H)^0$ *be a weighted shift (with positive weights) which attains its norm. Then* $T \in ZUCA$.

Proof. Let m be an integer such that $\alpha_m = \|T\|$. Suppose that $0 \neq C \in \mathcal{P}(T)$ and $C \leftrightarrow T$. Then

$$\|T\|^2 = \|T - C\|^2 \geq \|(T - C)(e_m)\|^2 = \|\alpha_m e_{m+1} - C(e_m)\|^2.$$

By Remark 3, $k_0 \geq 3$, so, by Remark 4, e_{m+1} is orthogonal to $C(e_m)$. So $\|T\|^2 \geq \alpha_m^2 + \|C(e_m)\|^2$, whence $C(e_m) = 0$. Thus, by Remark 5, $C = 0$. This proves that $T \in ZUCA$.

Remark 6. We now use this proposition to prove that the inclusion $ZUC \cap \mathcal{C}(H)^0 \subset ZUCA$ is proper. Consider the operator

$$T = \sum_{n \text{ odd}} e_{n+1} \otimes \overline{e_n} + \sum_{n \text{ even}} \frac{1}{n} e_{n+1} \otimes \overline{e_n}.$$

From Proposition 4 and Theorem 4 it follows that $T \in \mathcal{C}(H)^0$ and $T \notin ZUC$. However, Proposition 5 is satisfied so $T \in ZUC$.

The condition, in Proposition 5, that T attain its norm may be relaxed to the condition that a subsequence of the weights approaches the norm relatively quickly. To be more precise, suppose for $T \notin ZUC$ we let k_0 be the smallest integer so that $k_0 > 1$ and the condition of Theorem 4 is satisfied. Then we have

Proposition 6. *If T is a weighted shift with positive weights, $T \in \mathcal{C}(H)^0$, $T \notin ZUC$, and for every $\beta > 0$ and $k \geq k_0$ there exists an m depending upon β and k so that*

$$\|T\|^2 - a_{m+1}^2 < (\beta a_k \cdots a_{m+k-1}^2)/(a_m \cdots a_1),$$

then $T \in ZUCA$.

The proof is omitted since its essence is contained in the proof of Proposition 5. This result enlarges the class of weighted shifts known to be in $ZUCA$.

3. A counterexample. It has been established that all normal operators with empty point spectrum and several other classes of hyponormal operators with empty point spectrum are in $ZUC \cap \mathcal{C}(H)^0 \subset ZUCA$. One might suspect that all hyponormal operators with empty point spectrum are in $ZUCA$. This is decidedly not the case as is demonstrated by the following proposition and its corollary.

Proposition 7. *There exists a quasinormal operator with empty point spectrum having a nonzero commuting compact operator.*

Proof. Let $H = l^2$, $\{e_n\}_{n=1}^\infty$ the standard orthonormal basis in l^2, and define $P_0(x) = \Sigma_{n=1}^\infty a_n x_n e_n$ where $x = \Sigma_{n=1}^\infty x_n e_n$ and $a_n > a_{n+1} > 0$ for all n. P_0 is a positive operator on l^2. Let $T = UP$ be the dilated shift operator defined by P_0, i.e., $\operatorname{dom}(T) = \bigoplus_1^\infty H_j$, $H_j = H$, $P = \bigoplus_1^\infty P_j$, $P_j = P_0$, and $U =$ unilateral shift on $\bigoplus_1^\infty H_j$. Now the point spectrum of T is empty since P_0 is injective, and $UP = PU$, so T is quasinormal.

We recall the Rellich criterion for compact operators: an operator C is compact if and only if for any $\epsilon > 0$ there exists a finite codimensional subspace V_ϵ such that $\|C|V_\epsilon\| \leq \epsilon$. Let $C_j \in \mathcal{C}(H)$. The Rellich criterion implies that $C = \bigoplus_1^\infty C_j$ defined on $\bigoplus_1^\infty H_j$ is compact if and only if $\|C_j\| \to 0$ as $j \to \infty$. It is also easy to verify that $C \leftrightarrow T$ if and only if

$$(*) \qquad P_0 C_j = C_{j+1} P_0 \quad \text{for all } j.$$

So it suffices to make a choice of C_j satisfying these equations and such that $\|C_j\| \to 0$.

Define $C_j = \sum_{n=1}^{\infty} \beta_n^{(j)} e_{n+1} \otimes \bar{e}_n$, where $\beta_n^{(1)} \downarrow 0$ as $n \to \infty$ and $\beta_{n+1}^{(j)} = \beta_{n+1}^{(j-1)} \alpha_{n+1}/\alpha_n$. Let us now require that $\sup_n (\alpha_{n+1}/\alpha_n) = A < 1$ (e.g., $\alpha_n = 2^{-(n+1)}$). Then $\beta_{n+1}^{(j)} \leq A^{j-1}\beta_{n+1}^{(1)} < A^{j-1}\beta_1^{(1)}$, and since $\|C_j\| = \sup_n \beta_n^{(j)}$ we have $\|C_j\| \to 0$ as $n \to \infty$. In addition, each C_j is compact since $\beta_n^{(j)} \to 0$ as $n \to \infty$. Finally, condition (*) is satisfied so that $C \leftrightarrow T$.

Corollary. *There exists a quasinormal operator with empty point spectrum having a nonzero commuting compact best approximant.*

Proof. Let T and C be as in the previous example, let N be a normal operator with empty point spectrum on some Hilbert space, and consider the operator $N \oplus T$, on the appropriate Hilbert space $\mathcal{H}$. $N \oplus T$ is a quasinormal operator, and its point spectrum is empty. Suppose that $\|N\| > \|T - C\|$ and $\|N\| > \|T\|$. In [7] it was proved that if C_1 is a best compact approximant to N and C_2 is a best compact approximant to T, then

$$\operatorname{dist}(N \oplus T, \mathcal{C}(\mathcal{H})) = \|N \oplus T - C_1 \oplus C_2\|.$$

Because 0 is a best compact approximant to N (by Proposition 2) and $\|T - C_2\| \leq \|T\| < \|N\|$, it follows that

$$\operatorname{dist}(N \oplus T, \mathcal{C}(\mathcal{H})) = \max\{\|N\|, \|T - C_2\|\} = \|N\|.$$

Thus if we let $K = 0 \oplus C \neq 0$, we see that $K \leftrightarrow N \oplus T$ and

$$\|N \oplus T - K\| = \|N\| = \operatorname{dist}(N \oplus T, \mathcal{C}(\mathcal{H})). \quad \text{Q.E.D.}$$

4. The discontinuous nature of *ZUC* (*ZUCA*). The relationship between the metric complement $\mathcal{C}(H)^0$ and its subsets *ZUCA* is interesting. For example, the possibility that *ZUCA* is dense in $\mathcal{C}(H)^0$ is an intriguing but open question. However, neither *ZUCA* nor *ZUC* is closed.

Proposition 8. *There is a sequence of selfadjoint operators with empty point spectrum that converges to the identity operator.*

Proof. Let S be any selfadjoint operator with empty point spectrum. Evidently $T_n = I + \epsilon_n S$, $\epsilon_n \to 0$, is a sequence of selfadjoint operators with empty point spectrum converging uniformly to I. Q.E.D.

By Theorem 1, the T_n's are in *ZUC* and *ZUCA*; however, I is in neither. Such a phenomenon illustrates the delicate and discontinuous nature of the *ZUCA* property since we have just exhibited a sequence of operators each of whose set of commuting best compact approximations is zero dimensional,

but whose (norm) limit has an infinite dimensional set of commuting best compact approximations.

REFERENCES

1. A. Brown, P. R. Halmos and A. L. Shields, *Cesàro operators*, Acta Sci. Math. (Szeged) 26 (1965), 125–137. MR 32 #4539.

2. L. A. Coburn, *Weyl's theorem for nonnormal operators*, Michigan Math. J. 13 (1966), 285–288. MR 34 #1846.

3. J. Deddens and T. Wong, *The commutant analytic Toeplitz operators*, Trans. Amer. Math. Soc. 184 (1973), 261–273. MR 48 #2819.

4. I. C. Gohberg and M. G. Kreĭn, *Introduction to the theory of linear nonselfadjoint operators in Hilbert space*, "Nauka", Moscow, 1965; English transl., Transl. Math. Monographs, vol. 18, Amer. Math. Soc., Providence, R. I., 1969. MR 36 #3137; 39 #7447.

5. P. R. Halmos, *A Hilbert space problem book*, Von Nostrand, Princeton, N. J., 1967. MR 34 #8178.

6. R. Holmes and B. Kripke, *Best approximation by compact operators*, Indiana Univ. Math. J. 21 (1971), 255–263. MR 45 #5718.

7. R. Holmes, B. Scranton and J. Ward, *Approximation from the space of compact operators and other M-ideals*, Duke Math. J. (to appear).

8. V. Lomonosov, *Invariant subspaces for the family of operators which commute with a completely continuous operator*, Funkcional. Anal. i Priložen. 7 (1974), 213–214. (Russian)

9. R. D. Nussbaum, *The radius of the essential spectrum*, Duke Math. J. 37 (1970), 473–478. MR 41 #9028.

10. A. L. Shields and L. J. Wallen, *The commutants of certain Hilbert space operators*, Indiana Univ. Math. J. 20 (1970/71), 777–788. MR 44 #4558.

11. E. C. Titchmarsh, *The theory of functions*, 2nd ed., Oxford Univ. Press, Oxford, 1939.

DIVISION OF MATHEMATICAL SCIENCES, PURDUE UNVERSITY, LAFAYETTE, INDIANA (Current address of R. B. Holmes)

Current address (B. E. Scranton): Daniel H. Wagner, Associates, Paoli, Pennsylvania 19301

Current address (J. D. Ward): Department of Mathematics, Texas A & M University, College Station, Texas 77843

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

ON SEMISIMPLE COMMUTATIVE SEMIGROUPS

BY

B. D. ARENDT[1]

ABSTRACT. This paper presents an application of radical theory to the structure of commutative semigroups via their semilattice decomposition. Maximal group congruences and semisimplicity are characterized for certain classes of commutative semigroups and N-semigroups.

The concepts of a radical theory and semisimplicity in semigroups analogous to that of ring theory have been studied by a number of authors, both as a general theory, and applied to specific classes of semigroups. (See e.g. [1], [3], [5]–[9].) In this paper we apply these techniques to a study of the structure of commutative semigroups.

Every semigroup S has a least congruence μ such that S/μ is a semilattice Y. Each congruence class of μ is a subsemigroup of S, and the collection $\{S_\alpha\}$, $\alpha \in Y$, of congruence classes is called the greatest *semilattice decomposition* of S. Conversely, if Y is a semilattice and $\{S_\alpha\}$, $\alpha \in Y$, is a collection of pairwise disjoint semigroups, then any semigroup $S = \bigcup\{S_\alpha: \alpha \in Y\}$ with the property that $S_\alpha S_\beta \subseteq S_{\alpha\beta}$ for $\alpha, \beta \in Y$ is a *semilattice composition* of the S_α. Thus a natural approach to the structure of a given type of semigroup is through a characterization of the semilattice decompositions, the structure of the S_α, and the semilattice compositions. The semilattice decomposition of commutative semigroups was described by Tamura and Kimura [12]. In this case the semigroups S_α are the maximal archimedean subsemigroups, where S_α archimedean means given any two elements of S_α each divides some power of the other. Conversely, the general solution of the semilattice composition problem is known (see [10, Theorem III.7.2]).

If τ is a congruence on a semigroup S, we say τ is *modular* if there exists an element e in S such that $(ex)\tau(xe)\tau x$ for all x in S. The *radical*

Received by the editors June 28, 1974.

AMS (MOS) subject classifications (1970). Primary 20M10.

Key words and phrases. Commutative semigroup, maximal modular congruence, radical, semisimple, N-semigroup, semilattice decomposition.

[1] This research was supported in part by a University of Missouri Summer Research Grant.

congruence ρ is defined to be the intersection of all the maximal, modular congruences on S, and S is said to be *semisimple* if $\rho = \iota$, the identity relation. We begin here a study of semisimple commutative semigroups. The semilattice decomposition of such semigroups is described, and we obtain partial results on the nature of the S_α and semilattice compositions. As a related problem, we also obtain conditions for the existence of maximal group congruences on certain classes of N-semigroups. My thanks to the referee for the suggestion of Theorem 3.

If S is a commutative semigroup, a mapping ϕ from S into S is a *translation* if $(xy)\phi = x(y\phi)$ for all x, y in S. We denote by $T(S)$ the set of all translations of S. If the commutative semigroup S is a semilattice of subsemigroups S_α and $\alpha > \beta$, then $S_\alpha \cup S_\beta$ is evidently an ideal extension of S_β by $S_\alpha \cup \{0\}$. For each $a \in S_\alpha$, let the mapping ϕ_a be defined on S_β by $x\phi_a = xa$ for all $x \in S_\beta$. Then $\phi_a \in T(S_\beta)$ and $\phi_{\alpha\beta}: a \to \phi_a$ is a homomorphism from S_α into $T(S_\beta)$. Further, every such homomorphism determines an extension $S_\alpha \cup S_\beta$ with multiplication defined in the obvious way [13].

Oehmke [9] has characterized maximal modular congruences on a commutative semigroup, and we state his result as a lemma.

Lemma 1. *Let S be a commutative semigroup and τ a maximal modular congruence on S. Then either*

(i) *S/τ is a cyclic group of prime order, or*

(ii) *S/τ is the semilattice $\{0, 1\}$.*

Theorem 2. *Let S be a commutative semisimple semigroup and let $S = \bigcup S_\alpha$, $\alpha \in Y$, be the greatest semilattice decomposition of S. Then each S_α is semisimple and cancellative, and $\alpha > \beta$ implies S_α is isomorphically embedded in $T(S_\beta)$.*

Proof. Let $x, y \in S_\alpha$, then S semisimple implies there exists a maximal, modular congruence τ on S such that $x \not\equiv y \pmod \tau$. The restriction τ_α of τ to S_α is clearly a congruence on S_α. Further, if τ is of type (ii) (Lemma 1) then τ_α cannot separate elements of S_α since S_α is archimedean and has no proper prime ideals. Thus τ and hence τ_α must be of type (i). It follows that τ_α is a maximal modular congruence on S_α, so S_α is semisimple. The fact that all the maximal, modular congruences on S_α are cancellative and their intersection is the identity congruence says S_α must be cancellative. Suppose $\alpha > \beta$ and $a, b \in S_\alpha$ with $a\phi_{\alpha\beta} = b\phi_{\alpha\beta}$, that is, $ca = cb$ for $c \in S_\beta$. If τ is any maximal, modular congruence on S of type (i) then obviously $(ca)\tau(cb)$ and thus $a\tau b$ since τ is cancellative. If τ is a type (ii) congruence

then again $a\tau b$ holds since a and b are in the same archimedean component. It follows that $a\tau b$ for all maximal, modular congruences τ on S, and by semisimplicity, $a = b$. Thus $\phi_{\alpha\beta}$ is one-to-one, proving the theorem.

Each S_α of the theorem is thus a commutative, cancellative, archimedean semigroup. If S_α contains an idempotent then it is an abelian group. If such a semigroup does not contain an idempotent it is called an *N-semigroup.* To complete a characterization of commutative semisimple semigroups it is now necessary to describe the semisimple archimedean components S_α, and then determine those semilattice compositions of the S_α which are semisimple. It is evident that an abelian group G is semisimple if and only if the Frattini subgroup of G is trivial, that is, $\bigcap\{G^p : p \text{ is a prime}\} = \{e\}$. If, in addition, G is periodic (torsion), then G is semisimple if and only if each of its p-primary components is elementary abelian. If G is finitely generated, it is semisimple if and only if its torsion subgroup is semisimple.

Theorem 3. *Let S be a semilattice of abelian groups $\{G_\alpha : \alpha \in Y\}$, with linking homomorphisms $\phi_{\alpha,\beta}$. Then the following are equivalent.*

(i) *S is semisimple.*

(ii) *Each $\phi_{\alpha,\beta}$, $\alpha \geq \beta$, is one-to-one and the direct limit $\varinjlim G_\alpha$ is semisimple.*

(iii) *S is the subdirect product of a semilattice and a semisimple group.*

Proof. We show (i) $\Leftrightarrow$ (iii) $\Leftrightarrow$ (ii). (i) $\Leftrightarrow$ (iii) is obvious.

Suppose (iii). Then, by Theorem 2, each $\phi_{\alpha,\beta}$ is 1-1. Further the maximum group homomorphic image S/σ is isomorphic to $\varinjlim G_\alpha$; here $\sigma = \{(a, b) \in S \times S : ea = eb \text{ for some } e^2 = e \in S\}$.

Suppose $S \subseteq E \times G$ where G is semisimple. Then the projection π of S onto G is a group homomorphism so that $\sigma \subseteq \pi \circ \pi^{-1}$. On the other hand, let $a = (a_1, a_2)$, $b = (b_1, b_2)$ be in S and suppose $a\pi = b\pi$. Then $a_2 = b_2$ and so $aa^{-1}bb^{-1}a = aa^{-1}bb^{-1}b$; that is $\pi \circ \pi^{-1} \subseteq \sigma$. Hence $\sigma = \pi \circ \pi^{-1}$ and $G \cong S/\sigma \cong \varinjlim G_\alpha$. Thus (ii) holds.

Assume (ii). Since each G_α is a group, $T(G_\alpha)$ is isomorphic to G_α, so for $\alpha > \beta$ we have G_α is embedded in G_β. Let a and b be distinct elements of S. If a and b are in different G_α then it is clear that there is a maximal modular congruence of type (ii) (Lemma 1) which separates a and b, so assume $a, b \in G_\alpha$. If e_β is the identity of G_β, then $ae_\beta = a\phi_{\alpha,\alpha\beta}$, and since the linking homomorphisms are one-to-one, $(a, b) \notin \sigma$, so that (iii) holds. S is in fact the subdirect product of Y and $\varinjlim G_\alpha$.

Similarly, one can obtain a global characterization of commutative semisimple semigroups as follows. A commutative semigroup S is semisimple if and only if it can be embedded in the direct product of a semilattice and a semisimple group.

In general the converse of Theorem 2 does not hold without additional assumptions, even in the case where each S_α is a group. Jordan has given a converse of Theorem 2 for H-semigroups that are inverse semigroups or periodic in [6, Theorem 3] and [7, Theorem 5], respectively, by assuming the existence of a collection of subsemigroups with certain properties. In each case the semisimple semigroups are necessarily semilattices of abelian groups, hence a special case of Theorem 2. For the periodic case the following corollary sharpens Theorem 5 of [7] by eliminating the extra condition to give a converse of Theorem 2.

Corollary 4. *Let S be a periodic semigroup which is a semilattice of semisimple abelian groups G_α, $\alpha \in Y$, such that the multiplication homomorphisms are one-to-one, then S is semisimple.*

Proof. Each G_α is periodic abelian and semisimple so its p-primary components are elementary abelian for each prime p. It follows that the p-primary components of $\varinjlim G_\alpha$ are elementary, so $\varinjlim G_\alpha$ is semisimple and hence so is S.

Another finiteness condition that yields a converse for a semilattice of groups is that the semilattice Y has a zero. (See [7, Corollary 5.5].)

Corollary 5. *Let the semigroup S be a semilattice of semisimple abelian groups G_α, $\alpha \in Y$, where Y has a zero, and such that the multiplication homomorphisms are one-to-one, then S is semisimple.*

Proof. In this case $\varinjlim G_\alpha$ is isomorphic to G_0 which is semisimple, and the semisimplicity of S follows.

We now turn our attention to the other possibility for the subsemigroups S_α, a commutative, cancellative, archimedean semigroup without idempotent, or N-semigroup. Examples of such semigroups are the positive integers, positive rationals and positive reals under addition. Tamura [11] has given the following characterization of N-semigroups. Let N denote the set of nonnegative integers and let G be any abelian group. Let I be a mapping from $G \times G$ to N satisfying:

(i) $I(a, b) = I(b, a)$, $a, b \in G$,

(ii) $I(a, b) + I(ab, c) = I(a, bc) + I(b, c)$, $a, b, c \in G$,

(iii) for $a \in G$, $l(a^m, a) > 0$ for some $m > 0$,

(iv) $l(e, e) = 1$ where e is the identity of G.

Denote by (G, l) the set $N \times G$ with the binary operation $(m, a)(n, b) = (m + n + l(a, b), ab)$. Then (G, l) is an N-semigroup, and every N-semigroup is obtained in this manner. We note that neither G nor l is uniquely determined by S.

If $S = (G, l)$ is an N-semigroup then its maximal modular congruences must all be of type (i) since it is archimedean. Hall [4] has characterized homomorphisms of an N-semigroup into an abelian group G' as follows. Let ϕ be any mapping from G into G' satisfying

$$(a\phi)(b\phi) = (e\phi)^{l(a,b)}(ab)\phi, \qquad a, b \in G. \tag{1}$$

Define θ_ϕ from S into G' by

$$(m, a)\theta_\phi = (e\phi)^m(a\phi). \tag{2}$$

Then θ_ϕ is a homomorphism of S into G', and every homomorphism is of this type. Thus, to characterize maximal congruences on N-semigroups, and thereby the semisimple N-semigroups, we need to determine the existence of mappings satisfying (1) onto a cyclic group of prime order. In general this is difficult without additional assumptions on either G or l. Note that a mapping ϕ satisfying (1) will be a group homomorphism if and only if $e\phi = e'$, the identity of G'. Otherwise, $e\phi$ will be a generator of G' since it is of prime order. It is the latter case that is our primary interest since the homomorphism theory is well known. For $S = (G, l)$ and $a \in G$ with $|a| = m$, the order of a, we denote $l(a) = \sum_{k=1}^m l(a, a^k)$.

Theorem 6. *Let $S = (G, l)$ with G generated by a and let G' be a group of order p, a prime. There exists a mapping ϕ from G onto G' satisfying* (1) *which is not a homomorphism except for the case where G is finite, p divides $|G|$, and p does not divide $l(a)$. The value of $e\phi$ in G' is arbitrary ($\neq e'$). If G is finite and p does not divide $|G|$ then $a\phi$ is uniquely determined by $e\phi$, otherwise it may also be chosen arbitrarily.*

Proof. If G is infinite cyclic, then let $e\phi = g \neq e'$ in G' and let $a\phi = g^i$ for some $1 \leq i \leq p$. Inductively we define for $n > 1$,

$$a^n\phi = g^{k(n)}, \quad \text{where } k(n) = ni - \sum_{j=1}^{n-1} l(a, a^j), \tag{3}$$

and $a^{-1}\phi = g^k$, where $k \equiv l(a, a^{-1}) - i + 1 \pmod p$ and hence is uniquely

determined modulo p given i. If G has order n, then using (3) we see that $a^n\phi = e\phi$ if and only if

$$ni \equiv I(a) \pmod{p}. \tag{4}$$

Obviously $(n, p) = 1$ implies a unique solution i, while $(n, p) = p = (I(a), p)$ implies i may be chosen arbitrarily. No solution exists in the single case of the hypotheses. To check (1) for any two elements a^k, a^j in G, the conditions on the exponents of g give equality if and only if

$$I(a^k, a^j) \equiv \sum_{r=1}^{k+j-1} I(a, a^r) - \sum_{r=1}^{k-1} I(a, a^r) - \sum_{r=1}^{j-1} I(a, a^r) \pmod{p}.$$

From Lemma 2 of [2],

$$I(a^k, a^j) = I(a, a^{k+j-1}) + \sum_{r=0}^{k-2} I(a, a^{j+r}) - \sum_{r=0}^{k-2} I(a, a^{k-1-r}),$$

which is easily seen to be equal to the right side of the congruence, so this congruence holds for all p, and (1) is satisfied if G is infinite or $j + k \leq n$. If $j + k > n$ then $j + k - n = t < n$ when G is finite of order n, and $a^{j+k} = a^t$. Then (4) gives

$$(j + k - t)i \equiv ni \equiv \sum_{r=1}^{n} I(a, a^r) \equiv \sum_{r=t}^{j+k-t-1} I(a, a^r)\ ,$$

so that $a^{j+k}\phi = a^t\phi$ and (1) holds in this case as well.

The next result allows us to extend our definition of ϕ to a large class of groups.

Theorem 7. *Let $S = (G, I)$ where G is the direct product $H \times K$. Assume ϕ_1 and ϕ_2 are defined on H and K respectively to satisfy (1) such that $e\phi_1 = e\phi_2$. Set $e\phi = e\phi_1$ and for $hk \in G$ define $(hk)\phi = (e\phi)^{-I(h,k)}(h\phi_1)(k\phi_2)$, then ϕ satisfies (1) on G.*

Proof. Let $x = ab$ and $y = cd$ be elements of G, where $a, c \in H$ and $b, d \in K$. Then

$$(x\phi)(y\phi) = (e\phi)^{-I(a,b)-I(c,d)}(a\phi_1)(b\phi_2)(c\phi_1)(d\phi_2)$$

and

$$(xy)\phi = (acbd)\phi = (e\phi)^{-I(ac,bd)}(ac)\phi_1(bd)\phi_2$$

$$= (e\phi)^{-I(ac,bd)-I(a,c)-I(b,d)}(a\phi_1)(c\phi_1)(b\phi_2)(d\phi_2).$$

Multiplying $(xy)\phi$ by $(e\phi)^{I(x,y)}$ makes the exponent of $e\phi$ equal to $I(ab, cd) - I(ac, bd) - I(a, c) - I(b, d)$. By Lemma 6 of [2], this is equal to $-I(a, b) - I(c, d)$, so (1) is satisfied.

If G is semisimple then there are lots of maximal congruences on G and it is natural to expect a relationship to the semisimplicity of $S = (G, I)$. The next result shows that this is the case at least when G is finitely generated.

Theorem 8. *Let $S = (G, I)$ where G is a finitely generated semisimple abelian group, then S is semisimple.*

Proof. If (m, a) and (n, b) are two elements of S with $a \neq b$ then any homomorphism ϕ separating a and b will give a homomorphism θ_ϕ separating (m, a) and (n, b). Thus consider $(m, a) \neq (n, a)$ in S and write $G = \Pi_{k=1}^{t} G_k$ where $G_k = \langle a_k \rangle$. Let p be a prime such that $(p, m - n) = (p, |G_k|) = (p, I(a_k)) = 1$ for $k = 1, \ldots, t$. By Theorems 6 and 7 we can define a mapping ϕ which is not a homomorphism from G onto a cyclic group G' of order p. Further, $(m, a)\theta_\phi \neq (n, a)\theta_\phi$ since $(p, m - n) = 1$, so S is semisimple.

If G is any abelian group and we define $I(a, b) = 1$ for all a, b in G, then $S = (G, I)$ is an N-semigroup. These N-semigroups are of interest since they give lots of examples, are relatively easy to study, and they turn out to be fundamental in a sense to the general theory. For this class we are able to obtain a converse to Theorem 8, though it is not true in general.

Theorem 9. *Let $S = (G, I)$ be an N-semigroup with G finitely generated or periodic and $I(a, b) = 1$ for all a, b in G. If S is semisimple, then G is semisimple.*

Proof. The torsion subgroup of G is the direct product of p-primary subgroups G_p. Let $a \in G_p$ with $|a| = p^n$, $n \geq 1$. If $n > 1$ then $a \neq a^{p+1}$ so there is a homomorphism θ_ϕ onto a group G' of prime order q separating $(0, a)$ and $(0, a^{p+1})$. Thus $a\phi = (0, a)\theta_\phi \neq (0, a^{p+1})\theta_\phi = a^{p+1}\phi$ so that ϕ is not a homomorphism. Since (1) holds, $e\phi = g$ is not the identity of G' and must therefore generate G'. Letting $a\phi = g^i$, $1 \leq i \leq q$, we have $a^k\phi = g^{ki-k+1}$ for all $k \in N$ by induction. If $q \neq p$ then $g = e\phi = a^{p^n}\phi = g^{p^n i - p^n + 1}$ implies $p^n i - p^n + 1 \equiv 1 \pmod q$ so $i \equiv 1 \pmod q$. But then $i = 1$ so $a\phi = a^{p+1}\phi$, a contradiction. On the other hand, if $q = p$, then $i \equiv (p+1)i - (p+1) + 1 \pmod p$ gives $a\phi = a^{p+1}\phi$ again. We conclude that $n = 1$, so G_p is elementary abelian and G is semisimple.

A general characterization of semisimple N-semigroups appears to be

quite complicated due to two factors. First is the complicated structure of G itself when it is not torsion or finitely generated. Secondly, the conditions on the function I are quite general so that large numbers of them exist and not much can be said about them [2]. For this reason, the conditions for semisimplicity turn out to be number theoretical restrictions on I. We conclude with a characterization for the two cases where G is finitely generated or $\sigma(p^\infty)$ for some prime p.

Theorem 10. *Let $S = (G, I)$ where G is finitely generated abelian with torsion subgroup $\Pi_{k=1}^s G_k$, $G_k = \langle a_k \rangle$ of order $r_k = p_k^{s_k}$ and $\Pi_{k=1}^s r_k = M$. For each k, let $M_k = M/r_k$ and $I_k = I(a_k)$. Then S is semisimple if and only if given any collection of integers l_k, $k = 1, \ldots, s$, satisfying $|l_k| \le s_k$ and $l_k \equiv 0 \pmod{p_k}$, if $\Sigma_{k=1}^s l_k M_k I_k = Mz$ for some $z \in N$, then there exist a prime p and some collection i_k satisfying $r_k i_k \equiv I_k \pmod p$, $k = 1, \ldots, s$, such that $z \not\equiv \Sigma_{k=1}^s l_k i_k \pmod p$.*

Proof. Let (m, x) and (n, y) be distinct elements of S. A homomorphism θ_ϕ of S onto a cyclic group of prime order where ϕ is also a homomorphism will separate (m, x) and (n, y) if and only if ϕ separates x and y. Thus if $x = \Pi a_k^{m_k} u$ and $y = \Pi a_k^{n_k} v$, where u, v are from the torsion free part of G, are not separated by a maximal group homomorphism, then we may assume $u = v$ and $p_k | (m_k - n_k)$, $k = 1, \ldots, s$. If ϕ satisfies (1) and is not a homomorphism then $e\phi = g \ne e'$ and $a_k \phi = g^{i_k}$ where i_k satisfies $r_k i_k \equiv I_k \pmod p$ by (4). From (2) and (3) we get $(m, x)\theta_\phi$ is the element g raised to the power $m + \Sigma_{k=1}^s m_k i_k + F_x$ where F_x is an integer depending only on x. Thus $(m, x)\theta_\phi = (n, y)\theta_\phi$ if and only if

$$(5) \qquad m - n \equiv \sum_{k=1}^{s} (n_k - m_k) i_k + F_y - F_x \pmod p.$$

Set $l_k = m_k - n_k$ and assume $M \nmid \Sigma_{k=1}^s l_k M_k I_k$. The terms of the summation depend only on x and y, so for any m, n we must have $M(m - n) \ne \Sigma l_k M_k I_k + M(F_y - F_x)$. Thus we can choose a prime $p \nmid M$ so that the conditions of Theorem 6 hold for each G_k, there are unique solutions i_k to $r_k i_k \equiv I_k \pmod p$, and $M(m - n) \not\equiv \Sigma l_k M_k I_k + M(F_y - F_x) \pmod p$. $M = r_k M_k$ and $I_k \equiv r_k i_k \pmod p$ imply $M(m - n) \not\equiv \Sigma M l_k i_k + M(F_y - F_x) \pmod p$, and since $(M, p) = 1$, (5) cannot hold. If on the other hand, $\Sigma l_k M_k I_k = Mz$ for some

integer z then M divides $\Sigma l_k M_k I_k + M(F_y - F_x)$. The above argument will work to give separation of (m, x) and (n, y) except when m and n satisfy $m - n = z + F_y - F_x$. Now (5) fails for such a pair m, n if and only if $z \not\equiv \Sigma l_k i_k \pmod p$ which is the condition of the theorem.

Theorem 11. *Let $G = \sigma(p^\infty)$ for some prime p and $S = (G, I)$ for some index function I. $G = \langle a_1, a_2, a_3, \ldots \rangle$, where $a_1^p = e$ and $a_n^p = a_{n-1}$ for $n > 1$. Let $I_n = I(a_n)$. Then S maps homomorphically onto a group of order q for every prime $q \neq p$. Further, S maps onto a group of order p if and only if p divides I_1.*

Proof. If $q \neq p$ and $G_n = \langle a_n \rangle$ then Theorem 6 gives a unique i_n satisfying (4) and we define $a_n \phi = g^{i_n}$ where $\langle g \rangle = G'$ of order q. This is well defined on G if and only if $a_n \phi = (a_{n+1}^p)\phi$. Using (3) and letting $K_n = \Sigma_{j=1}^{p-1} I(a_{n+1}, a_{n+1}^j)$, we get equality if and only if $i_n \equiv p i_{n+1} - K_n \pmod q$ for all n. Since $p \neq q$, this holds if and only if

$$p^n i_n \equiv p^{n+1} i_{n+1} - p^n K_n \pmod q \Leftrightarrow I_n \equiv I_{n+1} - p^n K_n \pmod q.$$

Substituting $a_n = a_{n+1}^p$ in I_n, the last congruence is

$$\sum_{j=1}^{p^n} I(a_{n+1}^p, a_{n+1}^{pj}) \equiv I_{n+1} - p^n K_n \pmod q.$$

We now claim that we actually have equality in this last expression, that is, simplifying the notation a little, if $a \in G$ has order p^{n+1}, then

$$\sum_{j=1}^{p^n} I(a^p, a^{pj}) = \sum_{j=1}^{p^{n+1}} I(a, a^j) - p^n \sum_{j=1}^{p-1} I(a, a^j).$$

From Lemma 2 of [2] we have

$$I(a^p, a^{pj}) = I(a, a^{p+pj-1}) + \sum_{k=0}^{p-2} I(a, a^{pj+k}) - \sum_{k=0}^{p-2} I(a, a^{p-1-k}).$$

Simplifying gives

$$\sum_{j=1}^{p^n} I(a^p, a^{pj}) = \sum_{j=1}^{p^n} \sum_{k=0}^{p-1} I(a, a^{pj+k}) - p^n \sum_{k=1}^{p-1} I(a, a^k).$$

Using the fact that $a^{p^{n+1}} = e$, the first expression on the right becomes

$\sum_{j=1}^{pn+1} I(a, a^j)$ and our claim is proved. This implies ϕ is well defined and so a homomorphism onto G' exists. In particular we note that the above equality implies

$$I_n = I_{n+1} - p^n K_n \quad \text{for all } n. \tag{6}$$

If $q = p$ then Theorem 6 requires $I_n \equiv 0 \pmod p$ for all n and i_n is arbitrary. However ϕ is well defined if and only if $i_n \equiv - K_n \pmod p$ which has a unique solution for each n, giving the mapping θ_ϕ and conversely. Finally, (6) says $I_n \equiv 0 \pmod p$ for all n if and only if p divides I_1.

Corollary 12. *p^n divides I_n if and only if p^n divides I_{n+1} for all positive integers n.*

Theorem 13. *Let $G = \sigma(p^\infty)$ for some prime p and $S = (G, I)$ for some index function I. $G = \langle a_1, a_2, \ldots \rangle$ where $a_1^p = e$ and $a_n^p = a_{n-1}$ for $n > 1$. Let $I_n = I(a_n)$. Then S is semisimple if and only if for each positive integer n, given u such that $1 \leq u < p^n$, if p^n divides uI_n, then p^{n+1} does not divide uI_{n+1}.*

Proof. Let (t, x), (s, y) be distinct elements of S, then for some minimal n we have $x = a_n^j$, $y = a_n^k$, where (say) $j \geq k$. If $j = k$ then $t \neq s$ and to separate these two elements by (3) we need only satisfy $t - s \not\equiv 0 \pmod q$ which is possible by Theorem 11. Thus assume $j \neq k$. Again using (3), we are able to separate (t, x) and (s, y) if and only if there exists a prime q such that

$$(j - k)i_n \not\equiv t - s + \sum_{r=k}^{j-1} I(a_n, a_n^r) \pmod q. \tag{7}$$

If p^n does not divide $(j - k)I_n$, then $(j - k)I_n \neq p^n(t - s) + p^n \sum_{r=k}^{j-1} I(a_n, a_n^r)$ for any t and s, so there exists a prime $q \neq p$ such that

$$(j - k)I_n \not\equiv p^n(t - s) + p^n \sum_{r=k}^{j-1} I(a_n, a_n^r) \pmod q.$$

Choosing i_n to satisfy (4) gives

$$(j - k)p^n i_n \not\equiv p^n(t - s) + p^n \sum_{r=k}^{j-1} I(a_n, a_n^r) \pmod q,$$

and since $q \neq p$, (7) holds. If, on the other hand, $(j - k)I_n = p^n z$ for some integer z, then we obtain separation as above except for those integers t, s

such that $(j-k)I_n = p^n(t-s) + p^n \sum_{r=k}^{j-1} l(a_n, a_n^r)$, that is, $z = t - s + \sum_{r=k}^{j-1} l(a_n, a_n^r)$. It follows that we can separate (t, x) and (s, y) for such t, s if and only if there exist a prime q and i_n satisfying (4) such that $(j-k)i_n \not\equiv z \pmod q$. We show that q must equal p. Obviously $(j-k)I_n \equiv p^n z \pmod q$ for every prime q, and if $q \neq p$, then $z \equiv (p^n)^{-1}(j-k)I_n \pmod q$. If i_n satisfies (4), that is $p^n i_n \equiv I_n \pmod q$, then $(j-k)i_n \equiv (j-k)(p^n)^{-1}I_n \equiv z \pmod q$, so we must choose $q = p$. Since $j - k < p^n$ and $(j-k)I_n = p^n z$ we must have $p|I_n$ so S maps homomorphically onto a group of order p by Theorem 11. From the proof of Theorem 11 we observe $i_n \equiv -K_n \pmod p$, so we get separation if and only if $-(j-k)K_n \not\equiv z \pmod p$ or $z + (j-k)K_n \not\equiv 0 \pmod p$. Now $p^n z = (j-k)I_n = (j-k)I_{n+1} - (j-k)p^n K_n$ by (6) so $(j-k)I_{n+1} = p^n(z + (j-k)K_n)$. From this equation, $z + (j-k)K_n \equiv 0 \pmod p$ if and only if p^{n+1} divides $(j-k)I_{n+1}$, proving the theorem.

REFERENCES

1. B. D. Arendt, *Semisimple bands*, Trans. Amer. Math. Soc. 143 (1969), 133–143. MR 40 #255.
2. R. G. Biggs, M. Sasaki and T. Tamura, *Non-negative integer valued functions on commutative groups*. I, Proc. Japan Acad. 41 (1965), 564–569. MR 33 #2721.
3. A. H. Clifford, *Radicals in semigroups*, Semigroup Forum 1 (1970), no. 2, 103–127. MR 42 #1922.
4. R. E. Hall, *The translational hull of an N-semigroup*, Pacific J. Math. 41 (1972), 379–389.
5. H.-J. Hoehnke, *Structure of semigroups*, Canad. J. Math. 18 (1966), 449–491. MR 33 #5762.
6. M. J. Jordan, S.C., *Inverse H-semigroups, and t-semisimple inverse H-semigroups*, Trans. Amer. Math. Soc. 163 (1972), 75–84.
7. ———, *Periodic H-semigroups, and t-semisimple periodic H-semigroups*, Pacific J. Math. 41 (1972), 437–446. MR 46 #5498.
8. D. R. LaTorre, *Modular congruences and the Brown-McCoy radical for semigroups*, Proc. Amer. Math. Soc. 29 (1971), 427–433. MR 43 #6350.
9. R. H. Oehmke, *On maximal congruences and finite semisimple semigroups*, Trans. Amer. Math. Soc. 125 (1966), 223–237. MR 34 #2739.
10. M. Petrich, *Introduction to semigroups*, Merrill, Columbus, Ohio, 1973.
11. T. Tamura, *Commutative nonpotent archimedean semigroup with cancellation law*. I, J. Gakugei Tokushima Univ. 8 (1957), 5–11. MR 20 #3224.
12. T. Tamura and N. Kimura, *On decompositions of a commutative semigroup*, Kōdai Math. Sem. Rep. 4 (1954), 109–112. MR 16, 670.
13. R. Yoshida, *Ideal extensions of semigroups and compound semigroups*, Mem. Res. Inst. Sci. Eng. Ritumeiken Univ. 13 (1965), 1–8.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MISSOURI, COLUMBIA, MISSOURI 65201

TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 208, 1975

SIMILARITY OF QUADRATIC FORMS AND ISOMORPHISM OF THEIR FUNCTION FIELDS

BY

ADRIAN R. WADSWORTH[1]

ABSTRACT. This paper considers the question: Given anisotropic quadratic forms Q and Q' over a field K (char $K \neq 2$), if their function fields are isomorphic must Q and Q' be similar? It is proved that the answer is yes if Q is a Pfister form or the pure part of a Pfister form, or a 4-dimensional form. The argument for Pfister forms and their pure parts does not generalize because these are the only anisotropic forms which attain maximal Witt index over their function fields. To handle the 4-dimensional case the following theorem is proved: *If Q and Q' are two 4-dimensional forms over K with the same determinant d, then Q and Q' are similar over K iff they are similar over $K[\sqrt{d}]$.* The example of Pfister neighbors suggests that quadratic forms arguments are unlikely to settle the original question for other kinds of forms.

Let Q be a nonsingular quadratic form defined on an n-dimensional vector space V $(n \geq 3)$ over a field K (char $K \neq 2$), with diagonal representation $(a_0, a_1, \ldots, a_{n-1})$. The function field of Q, K_Q, is the quotient field of $K[x_1, \ldots, x_{n-1}]/(a_0 + a_1x_1^2 + \cdots + a_{n-1}x_{n-1}^2)$, i.e., $K(x_1, \ldots, x_{n-2})[\sqrt{-\alpha}]$, where $\alpha = a_{n-1}^{-1}(a_0 + a_1x_1^2 + \cdots + a_{n-2}x_{n-2}^2)$. The field K_Q is uniquely determined (up to K-isomorphism) by Q, independent of the choice of diagonal representation. Further, if Q' is another quadratic form which is isometric to Q, or even similar to Q (i.e., with a diagonal representation $(aa_0, aa_1, \ldots, aa_{n-1})$ for some $a \in K^*$), then $K_{Q'} \cong K_Q$.

We consider here the converse of the last sentence:

(†) If Q and Q' are two anisotropic forms over K, such that $K_{Q'} \cong K_Q$, must Q and Q' be similar?

We must confine attention exclusively to anisotropic forms when $\dim Q \geq 4$, since K_Q is purely transcendental over K iff Q is isotropic.

Received by the editors April 23, 1974.

AMS (MOS) subject classifications (1970). Primary 15A63; Secondary 10C05.

Key words and phrases. Similar quadratic forms, Pfister form, function field.

[1] Much of the work presented here was a part of my doctoral dissertation, which was prepared while I was an N.S.F. Graduate Fellow. I would like to thank my advisor, Irving Kaplansky, for suggesting this problem, and for many constructive conversations on it.

It was proved long ago by Witt [7], using the theory of algebraic function fields in one variable, that the answer to (†) is yes if the forms are three-dimensional. In §1 we will give an affirmative answer if Q is a Pfister form or the pure part of a Pfister form (in particular, reproving Witt's result). In §2 we will show that the answer is again yes if $\dim Q = 4$. We will indicate why the arguments used here do not generalize to other cases. Indeed, the example of Pfister neighbors suggests tha quadratic-form-theoretic arguments will not settle (†) except in the cases analyzed here.(2)

We begin with a few remarks on notation and terminology, and a standard lemma. Throughout the discussion the field K will be fixed, with char $K \neq 2$. Field isomorphisms will be K-isomorphisms. Each quadratic form Q will be nonsingular and (unless indicated otherwise) will be a K-form, i.e., a quadratic form defined on some finite-dimensional K-vector space V. As usual, $\dim Q$ is taken to be the dimension of V. The *discriminant* of Q, $\operatorname{disc} Q = \bar{d}$, is the image in $K^*/(K^*)^2$ of the determinant d of some matrix representing Q. (No sign is attached.) For Q' another K-form, $Q' \cong Q$ means Q and Q' are isometric. Q' is a *subform* of Q if there is another K-form Q'' such that $Q \cong Q' \perp Q''$. $Q \sim a$ means Q represents a. For $a_1, a_2, \ldots \in K^*$, $(a_1, \ldots, a_n)$ is an n-dimensional K-form which is represented by a diagonal matrix with entries $a_1, \ldots, a_n$. If $Q \cong (a_1, \ldots, a_n)$, then $aQ \cong (aa_1, \ldots, aa_n)$. Q is *nearly hyperbolic* if it is the orthogonal sum of a hyperbolic form and a 1-dimensional form. If L is a field containing K, Q_L denotes the L-form defined on $V \otimes_K L$ induced by Q on V. For standard quadratic form results quoted without reference, the reader is referred to [4] or [2].

A *Pfister form* P is a K-form representable as a product $P \cong \bigotimes_{j=1}^{r}(1, b_j)$, for some $b_1, \ldots, b_r \in K^*$. Recall that for any $a \in K^*$, if $P \sim a$, then $P \cong aP$ and if P is isotropic, then P is hyperbolic. Of course, P_L is again a Pfister form for any field $L \supseteq K$. Since $P \sim 1$ we have the decomposition $P \cong (1) \perp Q$. Q is called the *pure part* of the Pfister form P. Observe that Q_L is the pure part of P_L, and if Q_L is isotropic, then it is nearly hyperbolic. A *Pfister neighbor* (following Knebusch's definition) is a quadratic form R which is similar to a subform of a Pfister form P, with $\dim R > \frac{1}{2} \dim P$. For details on Pfister forms, see Pfister's original paper [6], or the excellent accounts in [2] and [3].

We state for the reader's convenience a well-known lemma which pro-

(2) All of the results given here for the function field of a form apply equally well to its homogeneous function field, which is a simple purely transcendental extension of the function field.

vides the opening wedge for our consideration of function fields. See, for example, [2, p. 200] for a proof.

Lemma 1. *Take $d \in K^*$ and let $L = K[\sqrt{d}]$. Any anisotropic K-form Q can be decomposed $Q \cong (1, -d) \otimes Q_0 \perp Q_1$, where Q_{1L} is anisotropic. (In particular, if Q_L is hyperbolic, the Q_1 term disappears.)*

1. Function fields of Pfister forms. The following notation will be used throughout this section. Q will be a K-form of dimension $n \geq 3$, $m = n - 2$. $F = K(x_1, \ldots, x_m)$ and $M = F(x_{m+1})$, where the x_i's are independent indeterminates. Take a diagonal representation $(a_0, \ldots, a_m, a)$ of Q, and let $\alpha = a^{-1}(a_0 + a_1x_1^2 + \cdots + a_mx_m^2)$ and $\beta = \alpha + x_{m+1}^2$. So $K_Q \cong F[\sqrt{-\alpha}]$.

Our first theorem collects the information needed to consider isomorphism of function fields of Pfister forms and their pure parts.

Theorem 2. *Suppose Q' is an anisotropic quadratic form over K. Assume further (with Q, α, β, as given above):*

(a) *Q' becomes hyperbolic in K_Q.*

Then,

(b) *there is an F-form Q_0, such that $Q'_F \cong (1, \alpha) \otimes Q_0$;*

(c) *$Q'_M \cong \beta Q'_M$;*

(d) *Q is similar to a subform of Q'. (Hence* dim $Q' \geq$ dim Q.*)*

Proof. We show (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d). Recall that the Witt index is preserved under purely transcendental field extensions. Thus, Q'_F is anisotropic. To obtain (b) from (a), apply Lemma 1 over F with $d = -\alpha$. Since $(1, \alpha)_M \sim \beta$, we have $(1, \alpha)_M \cong \beta(1, \alpha)_M$. (c) now follows at once. In particular, taking any $c \in K^*$ represented by Q', $Q'_M \sim c\beta$. Applying the Cassels-Pfister subform theorem [6] (inhomogeneous form, but the same proof holds), it follows that $(ca^{-1}a_0, \ldots, ca^{-1}a_m, c)$ is a subform of Q'. But this subform is just $ca^{-1}Q$, proving (d). Q.E.D.

Remark. (c) $\Rightarrow$ (a) holds as well. ((c) $\Rightarrow$ (b) can be deduced by the Cassels-Pfister subform theorem.) Indeed, the equivalence of (a) and (c) is a special case of a beautiful theorem of Knebusch [1].

Theorem 3. *If Q' is an anisotropic Pfister form* (dim $Q' \geq 4$) *and $K_Q \cong K_{Q'}$, then Q and Q' are similar.*

Proof. Being a Pfister form Q' becomes not merely isotropic, but actually hyperbolic in $K_{Q'}$. Since $K_Q \cong K_{Q'}$, Theorem 2 implies Q is similar to a subform of Q'. But dim Q = dim Q' because K_Q and $K_{Q'}$ have the same transcendence degree over K. So the subform must be all of Q', completing the proof.

Theorem 4. *If Q' is the pure part of an anisotropic Pfister form* ($\dim Q' \geq 3$) *and $K_Q \cong K_{Q'}$, then Q and Q' are similar.*

Proof. Say $P \cong (1) \perp Q'$, a Pfister form. Over K_Q, Q' becomes isotropic, so P becomes hyperbolic. By Theorem 2, Q is similar to a subform of P. But, by the transcendence degree argument just used, $\dim Q = \dim Q'$. Thus, $P \cong cQ \perp (d)$ for some $c, d \in K^*$. Since $P \sim d$, $P \cong dP \cong cdQ \perp (1)$. By the Witt cancellation theorem, $Q' \cong cdQ$. Q.E.D.

Note that the same kind of proof shows that if Q' is a Pfister neighbor of the anisotropic Pfister form P and $K_Q \cong K_{Q'}$, then Q is also a Pfister neighbor of P and $\dim Q = \dim Q'$. (But see the comments at the end of §2.)

Since there is only one similarity class of isotropic 3-dimensional forms, the condition that Q' be anisotropic can be deleted when $\dim Q' = 3$. Thus, we have reproved Witt's result which was the starting point of this investigation.

One might hope to generalize Theorems 3 and 4 to anisotropic forms which become hyperbolic or nearly hyperbolic over their function fields. The next two theorems show that there is no further generalization.(3)

Theorem 5. *If Q becomes hyperbolic over its function field* ($\dim Q \geq 4$), *then Q is already hyperbolic over K, or Q is similar to an anisotropic Pfister form.*

Proof. If Q is isotropic, then K_Q is purely transcendental over K. So Q hyperbolic over K_Q forces Q hyperbolic over K.

Now assume Q is anisotropic. Applying a suitable similarity factor, we may assume $Q \sim 1$. In the notation established above (with $a = 1$), we take $Q' = Q$, and Theorem 2 implies $Q_M \cong \beta Q_M$. That is, Q is strongly multiplicative in the sense of Pfister [6] (inhomogeneous form). Being anisotropic, Q must be a Pfister form. Q.E.D.

Theorem 6. *If Q becomes nearly hyperbolic over its function field* ($\dim Q \geq 3$), *then Q is already nearly hyperbolic over K, or Q is similar to the pure part of an anisotropic Pfister form.*

Proof. The isotropic case is handled just as in the preceding proof. So assume Q is anisotropic. We simplify the notation by applying a similarity to obtain $Q \sim 1$ (which permits us to take $a = 1$ in the expressions for α and β). Let $\bar{d} = \operatorname{disc} Q$.

Now, Q_F is anisotropic, but Q becomes nearly hyperbolic over $F[\sqrt{-\alpha}] = K_Q$. By Lemma 1, $Q_F \cong (1, \alpha) \otimes Q_0 \perp (\delta)$, for some F-form Q_0.

---

(3) Theorems 5 and 6 have been obtained independently by Knebusch.

(δ) can be determined by a comparison of discriminants. There are two possibilities, the first of which will be ruled out.

Case 1. $\dim Q \equiv 1 \pmod 4$. Then $\dim Q_0$ is even, so that $\operatorname{disc}((1, \alpha) \otimes Q_0) = \bar{1}$. Hence, we can take $\delta = d$, which shows that $Q_F \sim d$. However, as F is purely transcendental over K, Cassel's theorem [3, p. 18] implies $Q \sim d$. This gives a decomposition of Q over K: $Q \cong Q_1 \perp (d)$. Cancelling $(d)_F$ from the two decompositions of Q_F yields $Q_{1F} \cong (1, \alpha) \otimes Q_0$. By Theorem 2 ((b) $\Longrightarrow$ (d)), Q must be similar to a subform of Q_1, which is absurd, as $\dim Q_1 = \dim Q - 1$. Thus, Case 1 can never occur.

Case 2. $\dim Q \equiv 3 \pmod 4$. This time $\dim Q_0$ is odd, so that $\operatorname{disc}((1, \alpha) \otimes Q_0) = \bar{\alpha}$, and $(\delta) \cong (\alpha d)$. Let P be the K-form $Q \perp (d)$, i.e., $P \cong (a_0, \dots, a_m, 1, d)$. To complete the proof it suffices to show that P is an anisotropic Pfister form. (Then $P \cong dP \cong (1) \perp dQ$. So dQ is the pure part of P.)

The decomposition of Q_F yields $P_F \cong (1, \alpha) \otimes (Q_0 \perp (d))$. So, by Theorem 2 ((b) $\Longrightarrow$ (c)), $P_M \cong \beta P_M$. Hence, $P'_M \cong \beta P'_M$, where P' is the anisotropic part of P. By Theorem 2 ((c) $\Longrightarrow$ (d)), Q is similar to a subform of P'. Since $\dim Q = \dim P - 1$, P must be anisotropic.

Now take new independent indeterminates $y_0, \dots, y_n$ and set $K' = K(y_0, \dots, y_n)$ and $M' = M(y_0, \dots, y_n) = K'(x_1, \dots, x_{m+1})$. Of course, $P_{M'} \cong \beta P_{M'}$. Take any $c \in K'$, such that $P_{K'} \sim c$. Then $P_{M'} \sim c\beta$, so that, by the Cassels-Pfister subform theorem (treating $c\beta$ as a polynomial in the x_i's over K'), $(ca_0, ca_1, \dots, ca_m, c)$ is a subform of $P_{K'}$. A comparison of discriminants shows that $P_{K'} \cong (ca_0, ca_1, \dots, c, cd) \cong cP_{K'}$. In particular, we may take $c = a_0y_0^2 + a_1y_1^2 + \cdots + a_my_m^2 + y_{m+1}^2 + dy_n^2$, showing that P is strongly multiplicative, hence a Pfister form. Q.E.D.

2. 4-dimensional forms. We now proceed to give an affirmative answer to question (†) for 4-dimensional forms. This is done by using the Pfister form results over a quadratic extension L of K, then working back to K. The key step is the transition from L to K, which is provided by the following theorem.

Theorem 7. *Let Q and Q' be two 4-dimensional K-forms, each representing* 1 *and each with discriminant $\bar{d}$. Let $L = K[\sqrt{d}]$. Then Q and Q' are similar iff $Q_L \cong Q'_L$.*

Proof. *Necessity.* Q_L and Q'_L are Pfister forms, and hence isometric whenever they are similar.

Sufficiency. Suppose $Q_L \cong Q'_L$. Assume $d \notin (K^*)^2$—otherwise $L = K$ and there is nothing to prove. If Q is isotropic, then so are Q_L and Q'_L,

hence so is Q' (a well-known fact which can be deduced from Lemma 1 by a discriminant argument). But then disc Q = disc Q' implies that Q and Q' are similar. Thus, we may assume Q and Q' are anisotropic.

Take decompositions $Q \cong (1) \perp R$ and $Q' \cong (1) \perp R'$, and let $S = R \perp (-R')$. Since $Q_L \cong Q'_L$, S_L must be hyperbolic. If S were anisotropic, Lemma 1 would force disc $S = -\bar{d}$. But disc S = disc $R \cdot$ ($-$disc R') = $-\bar{1} \neq -\bar{d}$. Therefore, S must be isotropic. Since R and R' are anisotropic it follows that there is an $a \in K^*$, such that $R \sim a$ and $R' \sim a$. Hence, there exist diagonal representations $Q \cong (1, a, b, bad)$ and $Q' \cong (1, a, b', b'ad)$, for some $b, b' \in K^*$.

We will demonstrate the similarity of Q and Q' "piecewise", by finding a $t \in K^*$ with $(1, a) \cong t(1, a)$ and $(b', b'ad) \cong t(b, bad)$. To this end, consider the form $T = (1, a, -bb', -bb'ad)$, whose discriminant is $\bar{d}$. Working over L, where d is a square, we have

$$bT_L \cong (b, ba, -b', -b'ad)_L \cong (b, bad, -b', -bad)_L.$$

Thus, bT_L is a subform of the hyperbolic form $Q_L \perp -Q'_L$, with complement $(1, a, -1, -a)_L$ which is also hyperbolic. Therefore, T_L is hyperbolic. Using again the well-known fact quoted above, T itself is isotropic. From a nontrivial representation of 0 by T, we obtain $t \in K^*$, such that $(1, a) \sim t$ and $(bb', bb'ad) \sim t$. The two piecewise similarities now follow at once. Thus, $Q' \sim tQ$. Q.E.D.

This theorem may be of some interest in its own right. One immediate consequence is that a Hasse principle for similarity of 4-dimensional forms over a global field may be deduced from the Hasse principle for isometry.(4) (The global square theorem [4, 65:15] shows that equality of discriminants can be verified locally.)

For the proof of our final theorem, it is convenient to recast Theorem 7 as follows: *If Q and Q' are 4-dimensional forms each with discriminant $\bar{d}$ and $L = K[\sqrt{d}]$, then Q and Q' are similar iff Q_L and Q'_L are similar.*

Theorem 8. *Let Q and Q' be anisotropic 4-dimensional forms with $K_Q \cong K_{Q'}$. Then Q and Q' are similar.*

Proof. Let $\bar{d}$ = disc Q and $\overline{d'}$ = disc Q'. If $\bar{d} = \bar{1}$, then Q is similar to an anisotropic Pfister form (with the same function field), and Theorem 3 applies. Therefore, we may assume that $d, d' \notin (K^*)^2$.

Let $L = K[\sqrt{d}]$. Note that because K_Q is a regular extension of K, the K-isomorphism of K_Q and $K_{Q'}$ induces an L-isomorphism of L_{Q_L} of $L_{Q'_L}$.

(4) Ono [5] has shown that the Hasse principle for similarity holds for forms of any dimension over a global field.

By Theorem 3 (Q_L being similar to a Pfister form), Q_L and Q'_L are similar. In particular, disc Q_L = disc Q'_L, i.e., $d' \in (L^*)^2$. Since $d' \in K^* - (K^*)^2$, an easy computation (or an application of Lemma 1 to $\langle 1, -d' \rangle$) shows that $\overline{d'} = \overline{d}$. Thus, we may apply Theorem 7 (as recast) to conclude that Q and Q' are similar. Q.E.D.

The techniques used for Pfister forms and 4-dimensional forms do not generalize readily to other kinds of forms. To illustrate the problems that can be encountered, it suffices to consider the example of Pfister neighbors. Let Q and Q' be two *dissimilar* forms of the same dimension, which are each a Pfister neighbor of the same anisotropic Pfister form P. (So dim $P = 2^r$, for $r \geq 3$, and $2^{r-1} < \dim Q < 2^r - 1$.) Then in any extension field F of K, Q_F is isotropic iff P_F is hyperbolic iff Q'_F is isotropic. In particular, Q is isotropic over $K_{Q'}$, and vice versa. (In fact, it can be shown that for any K-form R, R is isotropic (resp. hyperbolic) over K_Q iff R is isotropic (resp. hyperbolic) over $K_{Q'}$.) Further, for every algebraic extension L of K, L_{Q_L} is purely transcendental over L iff $L_{Q'_L}$ is purely transcendental over L. Whether K_Q and $K_{Q'}$ can be isomorphic is unclear. It appears that new techniques will be necessary to settle this question.

Added in proof. M. Knebusch has shown that dissimilar Pfister neighbors often have isomorphic function fields. Specifically, if Q and Q' are neighbors of the Pfister form P, with dim $P = 2^r$, and if Q and Q' each have subforms similar to the Pfister form R, with $\dim R = 2^{r-1}$, and $\dim Q = \dim Q'$, then $K_Q \cong K_{Q'}$. This is one of many interesting results in his excellent paper *Generic splitting of quadratic forms*. I, to appear in Proc. London Math. Soc.

REFERENCES

1. M. Knebusch, *Specialization of quadratic and symmetric bilinear forms, and a norm theorem*, Acta Math. 24 (1973), 279–299.

2. T. Y. Lam, *The algebraic theory of quadratic forms*, Benjamin, Reading, Mass., 1973.

3. F. Lorenz, *Quadratische Formen über Körpern*, Lecture Notes in Math., vol. 130, Springer-Verlag, Berlin and New York, 1970. MR 44 #189.

4. O. T. O'Meara, *Introduction to quadratic forms*, Die Grundlehren der Math. Wissenschaften, Band 117, Academic Press, New York; Springer-Verlag, Berlin, 1963. MR 27 #2485.

5. T. Ono, *Arithmetic of orthogonal groups*, J. Math. Soc. Japan 7 (1955), 79–91. MR 16, 1087.

6. A. Pfister, *Multiplicative quadratische Formen*, Arch. Math. 16 (1965), 363–370. MR 32 #2408.

7. E. Witt, *Über ein Gegenbeispiel zum Normensatz*, Math. Z. 39 (1935), 462–467.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA 94720

Current address: Department of Mathematics, University of California at San Diego, La Jolla, California 92037

(*Continued from back cover*)

Submission of Manuscript

Mathematical papers intended for publication in the TRANSACTIONS or the MEMOIRS should be addressed to one of the editors. Subjects, and the editors associated with them, follow:

Real analysis (excluding harmonic analysis) and applied mathematics to FRANÇOIS TREVES, Department of Mathematics, Rutgers University, New Brunswick, N J 08903

Harmonic and complex analysis to HUGO ROSSI, Department of Mathematics, University of Utah, Salt Lake City, UT 84112

Abstract analysis to ALEXANDRA IONESCU TULCEA, Department of Mathematics, Northwestern University, Evanston, IL 60201

Algebra and number theory (excluding universal algebras) to STEPHEN S. SHATZ, Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19174

Logic, foundations, universal algebras and combinatorics to ALISTAIR H. LACHLAN, Department of Mathematics, Simon Fraser University, Burnaby 2, B.C., CANADA

Topology to PHILIP T. CHURCH, Department of Mathematics, Syracuse University, Syracuse, NY 13210

Global analysis and differential geometry to VICTOR W. GUILLEMIN, c/o Ms. M. McQuillin, Department of Mathematics, Harvard University, Cambridge, MA 02138

Probability and statistics to DANIEL W. STROOCK, Department of Mathematics, University of Colorado, Boulder, CO 80302

All other communications to the editors should be addressed to the Managing Editor, ALISTAIR H. LACHLAN.

CONTENTS

Vol. 208 1975 Whole No. 481

(Continued on inside back cover)